HANDBOOK OF ECONOMETRICS

Volume 5

James J. Heckman &
Edward Leamer

HANDBOOK OF ECONOMETRICS
VOLUME 5

HANDBOOKS IN ECONOMICS

2

Series Editors

KENNETH J. ARROW
MICHAEL D. INTRILIGATOR

ELSEVIER

AMSTERDAM • LONDON • NEW YORK • OXFORD • PARIS • SHANNON • TOKYO

HANDBOOK OF ECONOMETRICS

VOLUME 5

Edited by

JAMES J. HECKMAN
University of Chicago, Chicago

and

EDWARD LEAMER
University of California, Los Angeles

2001
ELSEVIER

AMSTERDAM • LONDON • NEW YORK • OXFORD • PARIS • SHANNON • TOKYO

ELSEVIER SCIENCE B.V.
Sara Burgerhartstraat 25
P.O. Box 211, 1000 AE Amsterdam, The Netherlands

First edition 2001

Library of Congress Cataloging in Publication Data
A catalog record from the Library of Congress has been applied for.

ISBN: 0-444-82340-9
ISSN: 0169-7218 (Handbooks of Economics Series)

♾ The paper used in this publication meets the requirements of ANSI/NISO Z39.48-1992 (Permanence of Paper).
Printed in The Netherlands.

INTRODUCTION TO THE SERIES

The aim of the *Handbooks in Economics* series is to produce Handbooks for various branches of economics, each of which is a definitive source, reference, and teaching supplement for use by professional researchers and advanced graduate students. Each Handbook provides self-contained surveys of the current state of a branch of economics in the form of chapters prepared by leading specialists on various aspects of this branch of economics. These surveys summarize not only received results but also newer developments, from recent journal articles and discussion papers. Some original material is also included, but the main goal is to provide comprehensive and accessible surveys. The Handbooks are intended to provide not only useful reference volumes for professional collections but also possible supplementary readings for advanced courses for graduate students in economics.

KENNETH J. ARROW and MICHAEL D. INTRILIGATOR

PUBLISHER'S NOTE

For a complete overview of the Handbooks in Economics Series, please refer to the listing at the end of this volume.

This Page Intentionally Left Blank

CONTENTS OF THE HANDBOOK

VOLUME 1

Part 1 – MATHEMATICAL AND STATISTICAL METHODS IN ECONOMETRICS

Chapter 1
Linear Algebra and Matrix Methods in Econometrics
HENRI THEIL

Chapter 2
Statistical Theory and Econometrics
ARNOLD ZELLNER

Part 2 – ECONOMETRIC MODELS

Chapter 3
Economic and Econometric Models
MICHAEL D. INTRILIGATOR

Chapter 4
Identification
CHENG HSIAO

Chapter 5
Model Choice and Specification Analysis
EDWARD E. LEAMER

Part 3 – ESTIMATION AND COMPUTATION

Chapter 6
Nonlinear Regression Models
TAKESHI AMEMIYA

Chapter 7
Specification and Estimation of Simultaneous Equation Models
JERRY A. HAUSMAN

Chapter 8
Exact Small Sample Theory in the Simultaneous Equations Model
PETER C. B. PHILLIPS

Chapter 9
Bayesian Analysis of Simultaneous Equation Systems
JACQUES H. DRÈZE and JEAN-FRANÇOIS RICHARD

Chapter 10
Biased Estimation
G. G. JUDGE and M. E. BOCK

Chapter 11
Estimation for Dirty Data and Flawed Models
WILLIAM S. KRASKER, EDWIN KUH, and ROY E. WELSCH

Chapter 12
Computational Problems and Methods
RICHARD E. QUANDT

VOLUME 2

Part 4 – TESTING

Chapter 13
Wald, Likelihood Ratio, and Lagrange Multiplier Tests in Econometrics
ROBERT F. ENGLE

Chapter 14
Multiple Hypothesis Testing
N. E. SAVIN

Chapter 15
Approximating the Distributions of Econometric Estimators and Test Statistics
THOMAS J. ROTHENBERG

Chapter 16
Monte Carlo Experimentation in Econometrics
DAVID F. HENDRY

Part 5 – TIME SERIES TOPICS

Chapter 17
Time Series and Spectral Methods in Econometrics
C. W. J. GRANGER and MARK W. WATSON

Chapter 18
Dynamic Specification
DAVID F. HENDRY, ADRIAN R. PAGAN, and J. DENIS SARGAN

Chapter 19
Inference and Causality in Economic Time Series Models
JOHN GEWEKE

Chapter 20
Continuous Time Stochastic Models and Issues of Aggregation over Time
A. R. BERGSTROM

Chapter 21
Random and Changing Coefficient Models
GREGORY C. CHOW

Chapter 22
Panel Data
GARY CHAMBERLAIN

Part 6 – SPECIAL TOPICS IN ECONOMETRICS: 1

Chapter 23
Latent Variable Models in Econometrics
DENNIS J. AIGNER, CHENG HSIAO, ARIE KAPTEYN, and TOM WANSBEEK

Chapter 24
Econometric Analysis of Qualitative Response Models
DANIEL L. McFADDEN

VOLUME 3

Part 7 – SPECIAL TOPICS IN ECONOMETRICS: 2

Chapter 25
Economic Data Issues
ZVI GRILICHES

Chapter 26
Functional Forms in Econometric Model Building
LAWRENCE J. LAU

Chapter 27
Limited Dependent Variables
PHOEBUS J. DHRYMES

Chapter 28
Disequilibrium, Self-selection, and Switching Models
G. S. MADDALA

Chapter 29
Econometric Analysis of Longitudinal Data
JAMES J. HECKMAN and BURTON SINGER

Part 8 – SELECTED APPLICATIONS AND USES OF ECONOMETRICS

Chapter 30
Demand Analysis
ANGUS DEATON

Chapter 31
Econometric Methods for Modeling Producer Behavior
DALE W. JORGENSON

Chapter 32
Labor Econometrics
JAMES J. HECKMAN and THOMAS E. MACURDY

Chapter 33
Evaluating the Predictive Accuracy of Models
RAY C. FAIR

Chapter 34
Econometric Approaches to Stabilization Policy in Stochastic Models of Macroeconomic Fluctuations
JOHN B. TAYLOR

Chapter 35
Economic Policy Formation: Theory and Implementation
(Applied Econometrics in the Public Sector)
LAWRENCE R. KLEIN

VOLUME 4

Part 9 – ECONOMETRIC THEORY

Chapter 36
Large Sample Estimation and Hypothesis Testing
WHITNEY K. NEWEY and DANIEL McFADDEN

Chapter 37
Empirical Process Methods in Econometrics
DONALD W. K. ANDREWS

Chapter 38
Applied Nonparametric Methods
WOLFGANG HÄRDLE and OLIVER LINTON

Chapter 39
Methodology and Theory for the Bootstrap
PETER HALL

Chapter 40
Classical Estimation Methods for LDV Models Using Simulation
VASSILIS A. HAJIVASSILOU and PAUL A. RUUD

Chapter 41
Estimation of Semiparametric Models
JAMES L. POWELL

Chapter 42
Restrictions of Economic Theory in Nonparametric Methods
ROSA L. MATZKIN

Chapter 43
Analog Estimation of Econometric Models
CHARLES F. MANSKI

Chapter 44
Testing Non-Nested Hypotheses
C. GOURIEROUX and A. MONFORT

Part 10 – THEORY AND METHODS FOR DEPENDENT PROCESSES

Chapter 45
Estimation and Inference for Dependent Processes
JEFFREY M. WOOLDRIDGE

Chapter 46
Unit Roots, Structural Breaks and Trends
JAMES H. STOCK

Chapter 47
Vector Autoregression and Cointegration
MARK W. WATSON

Chapter 48
Aspects of Modelling Nonlinear Time Series
TIMO TERÄSVIRTA, DAG TJØSTHEIM, and CLIVE W. J. GRANGER

Chapter 49
Arch Models
TIM BOLLERSLEV, ROBERT F. ENGLE, and DANIEL B. NELSON

Chapter 50
State-Space Models
JAMES D. HAMILTON

Chapter 51
Structural Estimation of Markov Decision Processes
JOHN RUST

VOLUME 5

Part 11 – NEW DEVELOPMENTS IN THEORETICAL ECONOMETRICS

Chapter 52
The Bootstrap
JOEL L. HOROWITZ

Chapter 53
Panel Data Models: Some Recent Developments
MANUEL ARELLANO and BO HONORÉ

Chapter 54
Interactions-based Models
WILLIAM A. BROCK and STEVEN N. DURLAUF

Chapter 55
Duration Models: Specification, Identification and Multiple Durations
GERARD J. VAN DEN BERG

Part 12 – COMPUTATIONAL METHODS IN ECONOMETRICS

Chapter 56
Computationally Intensive Methods for Integration in Econometrics
JOHN GEWEKE and MICHAEL KEANE

Chapter 57
Markov Chain Monte Carlo Methods: Computation and Inference
SIDDHARTHA CHIB

Part 13 – APPLIED ECONOMETRICS

Chapter 58
Calibration
CHRISTINA DAWKINS, T.N. SRINIVASAN, and JOHN WHALLEY

Chapter 59
Measurement Error in Survey Data
JOHN BOUND, CHARLES BROWN, and NANCY MATHIOWETZ

PREFACE TO THE HANDBOOK

The primary objective of Volume 5 of the Handbook of Econometrics and its companion Volume 6, is to collate in one place a body of research tools useful in applied econometrics and in empirical research in economics. A subsidiary objective is to update the essays on theoretical econometrics presented in the previous volumes of this series to include improvements in methods previously surveyed and methods not previously surveyed.

Part 11 contains four essays on developments in econometric theory. The essay by Joel Horowitz on the bootstrap presents a comprehensive survey of recent developments in econometrics and statistics on the application of the bootstrap to econometric models. With the decline in computing cost, bootstrapping offers an intellectually simpler alternative to the complex calculations required to produce asymptotic standard errors for complicated econometric models that sometimes displays better small properties than conventional estimators of standard errors. In applications, advice based on simple models is sometimes applied uncritically to the more complicated models estimated by economists. Horowitz provides a careful statement of conditions when the bootstrap works and when it fails that is of value to both theorists and empirical economists, and presents a variety of useful examples.

In the second essay, Manuel Arellano and Bo Honoré update the important essay by Gary Chamberlain on panel data in Volume 2 of this series to reflect developments in panel data methods in the past decade and a half. Their essay succinctly summarizes a large literature on using GMM methods to estimate panel data methods as well as the new work on nonlinear panel data methods developed by Honoré and his various coauthors.

In the third essay, William Brock and Steven Durlauf present the first rigorous econometric analysis of models of social interactions. This field has been an active area of research in economic theory and empirical work in the past decade but formal econometric analysis is scanty, although there are close parallels with the identification problems in this field and those in rational expectations econometrics. Indeed, the "reflection problem" discussed by Brock and Durlauf is just a version of the familiar problem of identification in self fulfilling equilibrium or rational expectations models [see e.g., Wallis (1980)]. Brock and Durlauf establish conditions under which models of social interactions can be identified and present constructive estimation strategies. They present a blueprint for future research in this rapidly growing area.

Gerard van den Berg's essay updates the essay by Heckman and Singer in Volume 3 of the Handbook to consider developments in the past decade and a half in econometric duration analysis. His essay presents a comprehensive discussion of multiple spell duration models which substantially extends the discussion in the published literature prior to this essay.

The essays in Part 12 present comprehensive surveys of new computational methods in econometrics. The advent of low cost computation has made many previously intractable econometric models empirically feasible, and has made Bayesian methods computationally attractive compared to classical methods. Bayesian methods replace optimization with integration and integration is cheap and numerically stable while optimization is neither. The essay by Geweke and Keane surveys a large literature in econometrics and statistics on computing integrals useful for Bayesian methods as well as in other settings. Chib focuses his essay on Markov Chain Monte Carlo Methods (MCMC) which have substantially reduced the cost of computing econometric models using Bayesian methods. This area has proven to be very fruitful and Chib summarizes the state of the art.

The essays on Applied Econometrics in Part 13 cover two main topics. The essay by Dawkins, Srinivasan and Whalley considers calibration as an econometric method. Calibration methods are widely used in applied general equilibrium theory and have been a source of great controversy in the econometrics literature. (See the symposium on calibration in the July, 1996 issue of the Journal of Economic Perspectives). Dawkins, Srinivasan and Whalley provide a careful account of current practice in calibrating applied general equilibrium models and the current state of the debate about the relative virtues of calibration vs. estimation.

The essay by Bound, Brown and Mathiowetz summarizes an impressive array of studies on measurement error and its consequences in economic data. Focusing primarily on data from labor markets, these authors document that the model of classical measurement error that has preoccupied the attention of econometricians for the past 50 years finds little support in the data. New patterns of measurement error are found that provide suggestions on what an empirically concordant model of measurement error would look like.

JAMES J. HECKMAN
University of Chicago, Chicago
EDWARD LEAMER
University of California, Los Angeles

References

Chamberlain, G. (1984), "Panel data", in: Z. Griliches and M. Intrilligator, eds., Handbook of Econometrics, Vol. II (North-Holland, Amsterdam) ch. 22.

Wallis, K. (1980), "Econometric implications of the rational expectations hypothesis", Econometrica XLVIII (1980):49–74.

CONTENTS OF VOLUME 5

Introduction to the Series v

Contents of the Handbook vii

Preface to the Handbook xiii

References xiv

Part 11 – NEW DEVELOPMENTS IN THEORETICAL ECONOMETRICS

Chapter 52
The Bootstrap
JOEL L. HOROWITZ 3159
Abstract 3160
Keywords 3160
1. Introduction 3161
2. The bootstrap sampling procedure and its consistency 3163
2.1. Consistency of the bootstrap 3166
2.2. Alternative resampling procedures 3169
3. Asymptotic refinements 3172
3.1. Bias reduction 3172
3.2. The distributions of statistics 3175
3.3. Bootstrap critical values for hypothesis tests 3179
3.4. Confidence intervals 3183
3.5. The importance of asymptotically pivotal statistics 3184
3.6. The parametric versus the nonparametric bootstrap 3185
3.7. Recentering 3186
4. Extensions 3188
4.1. Dependent data 3188
4.2. Kernel density and regression estimators 3195
4.3. Non-smooth estimators 3205
4.4. Bootstrap iteration 3210
4.5. Special problems 3212
4.6. The bootstrap when the null hypothesis is false 3212
5. Monte Carlo experiments 3213
5.1. The information-matrix test 3214

5.2. The t test in a heteroskedastic regression model 3215
5.3. The t test in a Box–Cox regression model 3217
5.4. Estimation of covariance structures 3218
6. Conclusions 3220
Acknowledgements 3221
Appendix A. Informal derivation of Equation (3.27) 3221
References 3223

Chapter 53
Panel Data Models: Some Recent Developments
MANUEL ARELLANO and BO HONORÉ 3229
Abstract 3231
Keywords 3231
1. Introduction 3232
2. Linear models with predetermined variables: identification 3233
2.1. Strict exogeneity, predeterminedness, and unobserved heterogeneity 3233
2.2. Time series models with error components 3241
2.3. Using stationarity restrictions 3248
2.4. Models with multiplicative effects 3250
3. Linear models with predetermined variables: estimation 3255
3.1. GMM estimation 3255
3.2. Efficient estimation under conditional mean independence 3259
3.3. Finite sample properties of GMM and alternative estimators 3262
3.4. Approximating the distributions of GMM and LIML for AR(1) models when the number of moments is large 3264
4. Nonlinear panel data models 3265
5. Conditional maximum likelihood estimation 3267
5.1. Conditional maximum likelihood estimation of logit models 3268
5.2. Poisson regression models 3269
6. Discrete choice models with “fixed” effects 3270
7. Tobit-type models with “fixed” effects 3272
7.1. Censored regression models 3272
7.2. Type 2 Tobit model (sample selection model) 3275
7.3. Other Tobit-type models 3276
7.4. Monotone transformation models 3278
7.5. Nonparametric regression and fixed effects 3280
7.6. Relationship with estimators for some cross sectional models 3280
8. Models with lagged dependent variables 3282
8.1. Discrete choice with state dependence 3283
8.2. Dynamic Tobit models 3286
8.3. Dynamic sample selection models 3287
9. “Random” effects models 3287

10. Concluding remarks 3290
References 3290

Chapter 54
Interactions-Based Models
WILLIAM A. BROCK and STEVEN N. DURLAUF 3297
Abstract 3299
Keywords 3299
1. Introduction 3300
  1.1. Neighborhoods and inequality 3301
  1.2. Spatial agglomeration 3302
  1.3. Technology choice 3302
  1.4. Preferences 3302
  1.5. Behavior of political parties 3303
  1.6. Social pathologies 3303
  1.7. Information cascades 3304
  1.8. Evolution of science 3304
  1.9. Chapter objectives 3304
2. Binary choice with social interactions 3305
  2.1. General framework 3305
  2.2. Global interactions 3309
  2.3. Local interactions 3312
  2.4. Relationship to statistical mechanics 3314
  2.5. Social planning problem 3315
  2.6. Linear-in-means model 3318
3. Identification: basic issues 3318
  3.1. Binary choice 3320
  3.2. Linear-in-means model 3322
  3.3. Instruments for unobservable expectations 3324
  3.4. Identification of individual versus neighborhood contextual effects 3325
  3.5. Nonlinear-in-means model 3326
  3.6. Implications of self-selection for identification 3328
  3.7. Implications of multiple equilibria for identification 3331
  3.8. Dynamic models and rational expectations 3331
4. Further topics in identification 3335
  4.1. Panel data 3335
  4.2. Duration data 3337
  4.3. Nonparametric approaches 3340
5. Sampling properties 3348
  5.1. Laws of large numbers 3348
  5.2. Naive estimator 3349
  5.3. Asymptotics for data generated by social planner 3349
  5.4. Unobserved variables 3350

6. Statistical analysis with grouped data 3352
6.1. Differences in cross-group behavior 3353
6.2. Spatial patterns 3353
6.3. Ecological inference 3354
7. Evidence 3355
7.1. Analyses under assumption of correct specification 3355
7.2. Analyses which are robust to unobserved correlated heterogeneity 3356
8. Summary and conclusions 3362
Appendix A 3363
A.1. Properties of binary choices made under a social planner 3363
A.2. Proof of Theorem 5 3368
A.3. Proof of Theorem 7 3369
References 3371

Chapter 55
Duration Models: Specification, Identification and Multiple Durations
GERARD J. VAN DEN BERG 3381
Abstract 3383
Keywords 3383
1. Introduction 3384
2. Basic concepts and notation 3387
3. Some structural models of durations 3389
3.1. Standard search model 3389
3.2. Repeated-search model 3393
4. The Mixed Proportional Hazard model 3394
4.1. Definition 3394
4.2. Time-varying explanatory variables 3397
4.3. Theoretical justification 3400
5. Identification of the MPH model with single-spell data 3405
5.1. Some implications of the MPH model specification 3405
5.2. Identification results 3408
5.3. Interaction between duration and explanatory variables in the observed hazard 3412
5.4. The sign of the interaction 3414
5.5. Specification of the unobserved heterogeneity distribution 3418
5.6. Effects of misspecification of functional forms 3422
6. The MPH model with multi-spell data 3426
6.1. Multi-spell data 3426
6.2. Identification results 3427
7. An informal classification of reduced-form multiple-duration models 3431
8. The Multivariate Mixed Proportional Hazard model 3437
8.1. Definition 3437
8.2. Identification results 3438
8.3. Specification of the bivariate unobserved heterogeneity distribution 3442

9. Causal duration effects and selectivity 3447
9.1. Lagged endogenous durations 3447
9.2. Endogenous shocks 3448
10. Conclusions and recommendations 3449
References 3453

Part 12 – COMPUTATIONAL METHODS IN ECONOMETRICS

Chapter 56
Computationally Intensive Methods for Integration in Econometrics
JOHN GEWEKE and MICHAEL KEANE 3463
Abstract 3465
Keywords 3465
1. Introduction 3466
2. Monte Carlo methods of integral approximation 3468
2.1. Independence sampling 3470
2.2. The Gibbs sampler 3472
2.3. The Hastings–Metropolis algorithm 3474
2.4. Some Markov chain Monte Carlo theory 3476
2.5. Metropolis within Gibbs 3479
2.6. Assessing numerical accuracy in Markov chain Monte Carlo 3479
3. Approximate solution of discrete dynamic optimization problems 3481
4. Classical simulation estimation of the multinomial probit model 3494
5. Univariate latent linear models 3506
5.1. An overview of the univariate latent linear model 3507
5.2. Some evidence from artificial data 3511
5.3. Some evidence from real data 3518
6. Multivariate latent linear models 3518
6.1. An overview of the multivariate latent linear model 3519
6.2. Some evidence from artificial data 3525
7. Bayesian inference for a dynamic discrete choice model 3538
Appendix A. The full univariate latent linear model 3548
Appendix B. The full multivariate latent linear model 3555
References 3564

Chapter 57
Markov Chain Monte Carlo Methods: Computation and Inference
SIDDHARTHA CHIB 3569
Abstract 3570
Keywords 3570
1. Introduction 3571
1.1. Organization 3573

2. Classical sampling methods 3573
2.1. Inverse transform method 3573
2.2. Accept–reject algorithm 3575
2.3. Method of composition 3576
3. Markov chains 3576
3.1. Definitions and results 3577
3.2. Computation of numerical accuracy and inefficiency factor 3579
4. Metropolis–Hastings algorithm 3580
4.1. The algorithm 3581
4.2. Convergence results 3584
4.3. Example 3585
4.4. Multiple-block M–H algorithm 3587
5. The Gibbs sampling algorithm 3589
5.1. The algorithm 3590
5.2. Connection with the multiple-block M–H algorithm 3591
5.3. Invariance of the Gibbs Markov chain 3592
5.4. Sufficient conditions for convergence 3592
5.5. Estimation of density ordinates 3592
5.6. Example: simulating a truncated multivariate normal 3594
6. Sampler performance and diagnostics 3595
7. Strategies for improving mixing 3596
7.1. Choice of blocking 3597
7.2. Tuning the proposal density 3597
7.3. Other strategies 3598
8. MCMC algorithms in Bayesian estimation 3599
8.1. Overview 3599
8.2. Notation and assumptions 3600
8.3. Normal and student-*t* regression models 3602
8.4. Binary and ordinal probit 3604
8.5. Tobit censored regression 3607
8.6. Regression with change point 3608
8.7. Autoregressive time series 3610
8.8. Hidden Markov models 3612
8.9. State space models 3614
8.10. Stochastic volatility model 3616
8.11. Gaussian panel data models 3619
8.12. Multivariate binary data models 3620
9. Sampling the predictive density 3623
10. MCMC methods in model choice problems 3626
10.1. Background 3626
10.2. Marginal likelihood computation 3627
10.3. Model space-parameter space MCMC algorithms 3633
10.4. Variable selection 3636

10.5. Remark 3639
11. MCMC methods in optimization problems 3639
12. Concluding remarks 3641
References 3642

Part 13 – APPLIED ECONOMETRICS

Chapter 58
Calibration

CHRISTINA DAWKINS, T.N. SRINIVASAN, and JOHN WHALLEY 3653
Abstract 3654
Keywords 3654
1. Introduction 3655
2. Calibration: its meaning and some early examples 3656
2.1. Two examples 3657
2.2. Reasons for using calibration 3660
3. The debate about calibration 3661
3.1. Lines of the debate 3662
3.2. Calibration is estimation, estimation is calibration 3663
3.3. The Jorgenson–McKitrick critique of micro model calibration 3665
4. Making calibration more concrete 3667
4.1. Calibrated macroeconomic models 3667
4.2. Calibrated microeconomic models 3671
4.3. Deficiencies of micro calibration 3677
4.4. Some recent examples of calibrations 3677
5. Best practice in calibration 3684
5.1. The choice of model for calibration 3684
5.2. The calibration of what to what? 3686
5.3. The choice of functional form 3687
5.4. The use of elasticity parameters 3689
6. New directions in calibration 3694
6.1. Expost decomposition and double or multiple calibration 3694
6.2. Sensitivity analysis 3695
6.3. Preadjusting data 3697
7. Conclusion 3698
References 3699

Chapter 59
Measurement Error in Survey Data

JOHN BOUND, CHARLES BROWN, and NANCY MATHIOWETZ 3705
Abstract 3707
Keywords 3707
1. Introduction 3708

2. The impact of measurement error on parameter estimates 3710
2.1. Special cases 3712
2.2. General results – linear model 3715
2.3. Differential measurement error – an example 3716
2.4. Bounding parameter estimates 3721
2.5. Contaminated and corrupted data 3723
2.6. Measurement error in categorical variables 3724
2.7. Nonlinear models 3727
3. Correcting for measurement error 3728
3.1. Instrumental variables in the bivariate linear model 3729
3.2. Multivariate linear model 3733
3.3. Nonlinear models 3735
3.4. The contribution of validation data 3737
4. Approaches to the assessment of measurement error 3740
5. Measurement error and memory: findings from household-based surveys 3743
5.1. Cognitive processes 3743
5.2. Social desirability 3745
5.3. Essential survey conditions 3746
5.4. Applicability of findings to the measurement of economic phenomena 3747
6. Evidence on measurement error in survey reports of labor-related phenomena 3748
6.1. Earnings 3748
6.2. Transfer program income 3770
6.3. Assets 3779
6.4. Hours worked 3784
6.5. Unemployment 3791
6.6. Industry and occupation 3802
6.7. Tenure, benefits, union coverage, size of establishment, and training 3805
6.8. Measurement error in household reports of health-related variables 3811
6.9. Education 3823
7. Conclusions 3830
References 3833

Author Index I-1

Subject Index I-19

Part 11

NEW DEVELOPMENTS IN THEORETICAL ECONOMETRICS

This Page Intentionally Left Blank

Chapter 52

THE BOOTSTRAP

JOEL L. HOROWITZ

Department of Economics, Northwestern University, Evanston, IL, USA

Contents

Abstract 3160
Keywords 3160
1. Introduction 3161
2. The bootstrap sampling procedure and its consistency 3163
  2.1. Consistency of the bootstrap 3166
  2.2. Alternative resampling procedures 3169
3. Asymptotic refinements 3172
  3.1. Bias reduction 3172
  3.2. The distributions of statistics 3175
  3.3. Bootstrap critical values for hypothesis tests 3179
  3.4. Confidence intervals 3183
  3.5. The importance of asymptotically pivotal statistics 3184
  3.6. The parametric versus the nonparametric bootstrap 3185
  3.7. Recentering 3186
4. Extensions 3188
  4.1. Dependent data 3188
    4.1.1. Methods for bootstrap sampling with dependent data 3188
    4.1.2. Asymptotic refinements in GMM estimation with dependent data 3191
    4.1.3. The bootstrap with non-stationary processes 3193
  4.2. Kernel density and regression estimators 3195
    4.2.1. Nonparametric density estimation 3196
    4.2.2. Asymptotic bias and methods for controlling it 3197
    4.2.3. Asymptotic refinements 3199
      4.2.3.1. The error made by first-order asymptotics when nh_n^{r+1} does not converge to 0 3201
    4.2.4. Kernel nonparametric mean regression 3202
  4.3. Non-smooth estimators 3205
    4.3.1. The LAD estimator for a linear median-regression model 3205
    4.3.2. The maximum score estimator for a binary-response model 3208
  4.4. Bootstrap iteration 3210
  4.5. Special problems 3212

Handbook of Econometrics, Volume 5, Edited by J.J. Heckman and E. Leamer

4.6. The bootstrap when the null hypothesis is false 3212
5. Monte Carlo experiments 3213
5.1. The information-matrix test 3214
5.2. The *t* test in a heteroskedastic regression model 3215
5.3. The *t* test in a Box–Cox regression model 3217
5.4. Estimation of covariance structures 3218
6. Conclusions 3220
Acknowledgements 3221
Appendix A. Informal derivation of Equation (3.27) 3221
References 3223

Abstract

The bootstrap is a method for estimating the distribution of an estimator or test statistic by resampling one's data or a model estimated from the data. Under conditions that hold in a wide variety of econometric applications, the bootstrap provides approximations to distributions of statistics, coverage probabilities of confidence intervals, and rejection probabilities of hypothesis tests that are more accurate than the approximations of first-order asymptotic distribution theory. The reductions in the differences between true and nominal coverage or rejection probabilities can be very large. The bootstrap is a practical technique that is ready for use in applications. This chapter explains and illustrates the usefulness and limitations of the bootstrap in contexts of interest in econometrics. The chapter outlines the theory of the bootstrap, provides numerical illustrations of its performance, and gives simple instructions on how to implement the bootstrap in applications. The presentation is informal and expository. Its aim is to provide an intuitive understanding of how the bootstrap works and a feeling for its practical value in econometrics.

Keywords

JEL classification: C12, C13, C15

1. Introduction

The bootstrap is a method for estimating the distribution of an estimator or test statistic by resampling one's data. It amounts to treating the data as if they were the population for the purpose of evaluating the distribution of interest. Under mild regularity conditions, the bootstrap yields an approximation to the distribution of an estimator or test statistic that is at least as accurate as the approximation obtained from first-order asymptotic theory. Thus, the bootstrap provides a way to substitute computation for mathematical analysis if calculating the asymptotic distribution of an estimator or statistic is difficult. The statistic developed by Härdle et al. (1991) for testing positive-definiteness of income-effect matrices, the conditional Kolmogorov test of Andrews (1997), Stute's (1997) specification test for parametric regression models, and certain functions of time-series data [Blanchard and Quah (1989), Runkle (1987), West (1990)] are examples in which evaluating the asymptotic distribution is difficult and bootstrapping has been used as an alternative.

In fact, the bootstrap is often more accurate in finite samples than first-order asymptotic approximations but does not entail the algebraic complexity of higher-order expansions. Thus, it can provide a practical method for improving upon first-order approximations. Such improvements are called *asymptotic refinements.* One use of the bootstrap's ability to provide asymptotic refinements is bias reduction. It is not unusual for an asymptotically unbiased estimator to have a large finite-sample bias. This bias may cause the estimator's finite-sample mean square error to greatly exceed the mean-square error implied by its asymptotic distribution. The bootstrap can be used to reduce the estimator's finite-sample bias and, thereby, its finite-sample mean-square error.

The bootstrap's ability to provide asymptotic refinements is also important in hypothesis testing. First-order asymptotic theory often gives poor approximations to the distributions of test statistics with the sample sizes available in applications. As a result, the nominal probability that a test based on an asymptotic critical value rejects a true null hypothesis can be very different from the true rejection probability (RP)[1]. The information matrix test of White (1982) is a well-known example of a test in which large finite-sample errors in the RP can occur when asymptotic critical values are used

[1] There is not general agreement on the name that should be given to the probability that a test rejects a true null hypothesis (that is, the probability of a Type I error). The source of the problem is that if the null hypothesis is composite, then the rejection probability can be different for different probability distributions in the null. Hall (1992a, p. 148) uses the word *level* to denote the rejection probability at the distribution that was, in fact, sampled. Beran (1988, p. 696) defines *level* to be the supremum of rejection probabilities over all distributions in the null hypothesis. Other authors [Lehmann (1959, p. 61); Rao (1973, p. 456)] use the word *size* for the supremum. Lehmann defines *level* as a number that exceeds the rejection probability at all distributions in the null hypothesis. In this chapter, the term *rejection probability* or *RP* will be used to mean the probability that a test rejects a true null hypothesis with whatever distribution generated the data. The RP of a test is the same as Hall's definition of level. The RP is different from the size of a test and from Beran's and Lehmann's definitions of *level.*

[Horowitz (1994), Kennan and Neumann (1988), Orme (1990), Taylor (1987)]. Other illustrations are given later in this chapter. The bootstrap often provides a tractable way to reduce or eliminate finite-sample errors in the RP's of statistical tests.

The problem of obtaining critical values for test statistics is closely related to that of obtaining confidence intervals. Accordingly, the bootstrap can also be used to obtain confidence intervals with reduced errors in coverage probabilities. That is, the difference between the true and nominal coverage probabilities is often lower when the bootstrap is used than when first-order asymptotic approximations are used to obtain a confidence interval.

The bootstrap has been the object of much research in statistics since its introduction by Efron (1979). The results of this research are synthesized in the books by Beran and Ducharme (1991), Davison and Hinkley (1997), Efron and Tibshirani (1993), Hall (1992a), Mammen (1992), and Shao and Tu (1995). Hall (1994), Horowitz (1997), Jeong and Maddala (1993) and Vinod (1993) provide reviews with an econometric orientation. This chapter covers a broader range of topics than do these reviews. Topics that are treated here but only briefly or not at all in the reviews include bootstrap consistency, subsampling, bias reduction, time-series models with unit roots, semiparametric and nonparametric models, and certain types of non-smooth models. Some of these topics are not treated in existing books on the bootstrap.

The purpose of this chapter is to explain and illustrate the usefulness and limitations of the bootstrap in contexts of interest in econometrics. Particular emphasis is given to the bootstrap's ability to improve upon first-order asymptotic approximations. The presentation is informal and expository. Its aim is to provide an intuitive understanding of how the bootstrap works and a feeling for its practical value in econometrics. The discussion in this chapter does not provide a mathematically detailed or rigorous treatment of the theory of the bootstrap. Such treatments are available in the books by Beran and Ducharme (1991) and Hall (1992a) as well as in journal articles that are cited later in this chapter.

It should be borne in mind throughout this chapter that although the bootstrap often provides smaller biases, smaller errors in the RP's of tests, and smaller errors in the coverage probabilities of confidence intervals than does first-order asymptotic theory, bootstrap bias estimates, RP's, and confidence intervals are, nonetheless, approximations and not exact. Although the accuracy of bootstrap approximations is often very high, this is not always the case. Even when theory indicates that it provides asymptotic refinements, the bootstrap's numerical performance may be poor. In some cases, the numerical accuracy of bootstrap approximations may be even worse than the accuracy of first-order asymptotic approximations. This is particularly likely to happen with estimators whose asymptotic covariance matrices are "nearly singular," as in instrumental-variables estimation with poorly correlated instruments and regressors. Thus, the bootstrap should not be used blindly or uncritically.

However, in the many cases where the bootstrap works well, it essentially removes getting the RP or coverage probability right as a factor in selecting a test statistic or method for constructing a confidence interval. In addition, the bootstrap can provide dramatic reductions in the finite-sample biases and mean-square errors of certain estimators.

The remainder of this chapter is divided into five sections. Section 2 explains the bootstrap sampling procedure and gives conditions under which the bootstrap distribution of a statistic is a consistent estimator of the statistic's asymptotic distribution. Section 3 explains when and why the bootstrap provides asymptotic refinements. This section concentrates on data that are simple random samples from a distribution and statistics that are either smooth functions of sample moments or can be approximated with asymptotically negligible error by such functions (the smooth function model). Section 4 extends the results of Section 3 to dependent data and statistics that do not satisfy the assumptions of the smooth function model. Section 5 presents Monte Carlo evidence on the numerical performance of the bootstrap in a variety of settings that are relevant to econometrics, and Section 6 presents concluding comments.

For applications-oriented readers who are in a hurry, the following list of bootstrap dos and don'ts summarizes the main practical conclusions of this chapter.

Bootstrap Dos and Don'ts

(1) **Do** use the bootstrap to estimate the probability distribution of an asymptotically pivotal statistic or the critical value of a test based on an asymptotically pivotal statistic whenever such a statistic is available. (Asymptotically pivotal statistics are defined in Section 2. Sections 3.2–3.5 explain why the bootstrap should be applied to asymptotically pivotal statistics.)

(2) **Don't** use the bootstrap to estimate the probability distribution of a non-asymptotically-pivotal statistic such as a regression slope coefficient or standard error if an asymptotically pivotal statistic is available.

(3) **Do** recenter the residuals of an overidentified model before applying the bootstrap to the model. (Section 3.7 explains why recentering is important and how to do it.)

(4) **Don't** apply the bootstrap to models for dependent data, semi- or nonparametric estimators, or non-smooth estimators without first reading Section 4 of this chapter.

2. The bootstrap sampling procedure and its consistency

The bootstrap is a method for estimating the distribution of a statistic or a feature of the distribution, such as a moment or a quantile. This section explains how the bootstrap is implemented in simple settings and gives conditions under which it provides a consistent estimator of a statistic's asymptotic distribution. This section also gives examples in which the consistency conditions are not satisfied and the bootstrap is inconsistent.

The estimation problem to be solved may be stated as follows. Let the data be a random sample of size n from a probability distribution whose cumulative distribution function (CDF) is F_0. Denote the data by $\{X_i: i=1,\ldots,n\}$. Let F_0 belong to a finite- or infinite-dimensional family of distribution functions, $\Im$. Let F denote a general member of $\Im$. If $\Im$ is a finite-dimensional family indexed by the parameter θ whose population value is θ_0, write $F_0(x, \theta_0)$ for $P(X \leqslant x)$ and $F(x, \theta)$ for a general member of the parametric family. Let $T_n = T_n(X_1,\ldots,X_n)$ be a statistic (that is, a function of the data). Let $G_n(\tau, F_0) \equiv P(T_n \leqslant \tau)$ denote the exact, finite-sample CDF of T_n. Let $G_n(\cdot, F)$ denote the exact CDF of T_n when the data are sampled from the distribution whose CDF is F. Usually, $G_n(\tau, F)$ is a different function of τ for different distributions F. An exception occurs if $G_n(\cdot, F)$ does not depend on F, in which case T_n is said to be *pivotal.* For example, the t statistic for testing a hypothesis about the mean of a normal population is independent of unknown population parameters and, therefore, is pivotal. The same is true of the t statistic for testing a hypothesis about a slope coefficient in a normal linear regression model. Pivotal statistics are not available in most econometric applications, however, especially without making strong distributional assumptions (e.g., the assumption that the random component of a linear regression model is normally distributed). Therefore, $G_n(\cdot, F)$ usually depends on F, and $G_n(\cdot, F_0)$ cannot be calculated if, as is usually the case in applications, F_0 is unknown. The bootstrap is a method for estimating $G_n(\cdot, F_0)$ or features of $G_n(\cdot, F_0)$ such as its quantiles when F_0 is unknown.

Asymptotic distribution theory is another method for estimating $G_n(\cdot, F_0)$. The asymptotic distributions of many econometric statistics are standard normal or chi-square, possibly after centering and normalization, regardless of the distribution from which the data were sampled. Such statistics are called *asymptotically pivotal,* meaning that their asymptotic distributions do not depend on unknown population parameters. Let $G_\infty(\cdot, F_0)$ denote the asymptotic distribution of T_n. Let $G_\infty(\cdot, F)$ denote the asymptotic CDF of T_n when the data are sampled from the distribution whose CDF is F. If T_n is asymptotically pivotal, then $G_\infty(\cdot, F) \equiv G_\infty(\cdot)$ does not depend on F. Therefore, if n is sufficiently large, $G_n(\cdot, F_0)$ can be estimated by $G_\infty(\cdot)$ without knowing F_0. This method for estimating $G_n(\cdot, F_0)$ is often easy to implement and is widely used. However, as was discussed in Section 1, $G_\infty(\cdot)$ can be a very poor approximation to $G_n(\cdot, F_0)$ with samples of the sizes encountered in applications.

Econometric parameter estimators usually are not asymptotically pivotal (that is, their asymptotic distributions usually depend on one or more unknown population parameters), but many are asymptotically normally distributed. If an estimator is asymptotically normally distributed, then its asymptotic distribution depends on at most two unknown parameters, the mean and the variance, that can often be estimated without great difficulty. The normal distribution with the estimated mean and variance can then be used to approximate the unknown $G_n(\cdot, F_0)$ if n is sufficiently large.

The bootstrap provides an alternative approximation to the finite-sample distribution of a statistic $T_n(X_1,\ldots,X_n)$. Whereas first-order asymptotic approximations replace the unknown distribution function G_n with the known function G_∞, the bootstrap replaces the unknown distribution function F_0 with a known estimator. Let F_n denote the estimator of F_0. Two possible choices of F_n are:

(1) The empirical distribution function (EDF) of the data:

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leqslant x),$$

where I is the indicator function. It follows from the Glivenko–Cantelli theorem that $F_n(x) \to F_0(x)$ as $n \to \infty$ uniformly over x almost surely.

(2) A parametric estimator of F_0. Suppose that $F_0(\cdot) = F(\cdot, \theta_0)$ for some finite-dimensional θ_0 that is estimated consistently by θ_n. If $F(\cdot, \theta)$ is a continuous function of θ in a neighborhood of θ_0, then $F(x, \theta_n) \to F(x, \theta_0)$ as $n \to \infty$ at each x. The convergence is in probability or almost sure according to whether $\theta_n \to \theta_0$ in probability or almost surely.

Other possible F_n's are discussed in Section 3.7.

Regardless of the choice of F_n, the bootstrap estimator of $G_n(\cdot, F_0)$ is $G_n(\cdot, F_n)$. Usually, $G_n(\cdot, F_n)$ cannot be evaluated analytically. It can, however, be estimated with arbitrary accuracy by carrying out a Monte Carlo simulation in which random samples are drawn from F_n. Thus, the bootstrap is usually implemented by Monte Carlo simulation. The Monte Carlo procedure for estimating $G_n(\tau, F_0)$ is as follows:

Monte Carlo Procedure for Bootstrap Estimation of $G_n(\tau, F_0)$

Step 1: Generate a bootstrap sample of size $n, \{X_i^*: i = 1, \ldots, n\}$, by sampling the distribution corresponding to F_n randomly. If F_n is the EDF of the estimation data set, then the bootstrap sample can be obtained by sampling the estimation data randomly with replacement.

Step 2: Compute $T_n^* \equiv T_n(X_1^*, \ldots, X_n^*)$.

Step 3: Use the results of many repetitions of steps 1 and 2 to compute the empirical probability of the event $T_n^* \leqslant \tau$ (that is, the proportion of repetitions in which this event occurs).

Procedures for using the bootstrap to compute other statistical objects are described in Sections 3.1 and 3.3. Brown (1999) and Hall (1992a, Appendix II) discuss simulation methods that take advantage of techniques for reducing sampling variation in Monte Carlo simulation. The essential characteristic of the bootstrap, however, is the use of F_n to approximate F_0 in $G_n(\cdot, F_0)$, not the method that is used to evaluate $G_n(\cdot, F_n)$.

Since F_n and F_0 are different functions, $G_n(\cdot, F_n)$ and $G_n(\cdot, F_0)$ are also different functions unless T_n is pivotal. Therefore, the bootstrap estimator $G_n(\cdot, F_n)$ is only an approximation to the exact finite-sample CDF of $T_n, G_n(\cdot, F_0)$. Section 3 discusses the accuracy of this approximation. The remainder of this section is concerned with conditions under which $G_n(\cdot, F_n)$ satisfies the minimal criterion for adequacy as an estimator of $G_n(\cdot, F_0)$, namely consistency. Roughly speaking, $G_n(\cdot, F_n)$ is consistent if it converges in probability to the asymptotic CDF of $T_n, G_\infty(\cdot, F_0)$, as $n \to \infty$. Section 2.1 defines consistency precisely and gives conditions under which it holds.

Section 2.2 describes some resampling procedures that can be used to estimate $G_n(\cdot, F_0)$ when the bootstrap is not consistent.

2.1. Consistency of the bootstrap

Suppose that F_n is a consistent estimator of F_0. This means that at each x in the support of X, $F_n(x) \to F_0(x)$ in probability or almost surely as $n \to \infty$. If F_0 is a continuous function, then it follows from Polya's theorem that $F_n \to F_0$ in probability or almost surely uniformly over x. Thus, F_n and F_0 are uniformly close to one another if n is large. If, in addition, $G_n(\tau, F)$ considered as a functional of F is continuous in an appropriate sense, it can be expected that $G_n(\tau, F_n)$ is close to $G_n(\tau, F_0)$ when n is large. On the other hand, if n is large, then $G_n(\cdot, F_0)$ is uniformly close to the asymptotic distribution $G_\infty(\cdot, F_0)$ if $G_\infty(\cdot, F_0)$ is continuous. This suggests that the bootstrap estimator $G_n(\cdot, F_n)$ and the asymptotic distribution function $G_\infty(\cdot, F_0)$ should be uniformly close if n is large and suitable continuity conditions hold. The definition of consistency of the bootstrap formalizes this idea in a way that takes account of the randomness of the function $G_n(\cdot, F_n)$. Let $\Im$ denote the space of permitted distribution functions.

Definition 2.1. Let P_n denote the joint probability distribution of the sample $\{X_i$: $i = 1, \dots, n\}$. The bootstrap estimator $G_n(\cdot, F_n)$ is consistent if for each $\epsilon > 0$ and $F_0 \in \Im$

$$\lim_{n \to \infty} P_n \left[\sup_\tau |G_n(\tau, F_n) - G_\infty(\tau, F_0)| > \varepsilon \right] = 0.$$

A theorem by Beran and Ducharme (1991) gives conditions under which the bootstrap estimator is consistent. This theorem is fundamental to understanding the bootstrap. Let ρ denote a metric on the space $\Im$ of permitted distribution functions.

Theorem 2.1 (Beran and Ducharme 1991). *$G_n(\cdot, F_n)$ is consistent if for any $\epsilon > 0$ and $F_0 \in \Im$: (i) $\lim_{n\to\infty} P_n[\rho(F_n, F_0) > \varepsilon] = 0$; (ii) $G_\infty(\tau, F)$ is a continuous function of τ for each $F \in \Im$; and (iii) for any τ and any sequence $\{H_n\} \in \Im$ such that $\lim_{n\to\infty} \rho(H_n, F_0) = 0$, $G_n(\tau, H_n) \to G_\infty(\tau, F_0)$.*

The following is an example in which the conditions of Theorem 2.1 are satisfied:

Example 2.1. *The distribution of the sample average*: Let $\Im$ be the set of distribution functions F corresponding to populations with finite variances. Let $\bar{X}$ be the average of the random sample $\{X_i$: $i = 1, \dots, n\}$. Define $T_n = n^{1/2}(\bar{X} - \mu)$, where $\mu = E(X)$. Let $G_n(\tau, F_0) = P_n\left[n^{1/2}(\bar{X} - \mu) \leqslant \tau\right]$. Consider using the bootstrap to estimate $G_n(\tau, F_0)$. Let F_n be the EDF of the data. Then the bootstrap analog of T_n is $T_n^* = n^{1/2}(\bar{X}^* - \bar{X})$, where $\bar{X}^*$ is the average of a random sample of size n drawn from F_n (the bootstrap sample). The bootstrap sample can be obtained by sampling the data $\{X_i\}$ randomly with replacement. T_n^* is centered at $\bar{X}$ because $\bar{X}$ is the mean of the distribution from which the bootstrap sample is drawn. The bootstrap estimator of $G_n(\tau, F_0)$ is $G_n(\tau, F_n) = P_n^* \left[n^{1/2}(\bar{X}^* - \bar{X}) \leqslant \tau\right]$, where P_n^* is the probability distribution induced by the bootstrap sampling process. $G_n(\tau, F_n)$ satisfies the conditions of Theorem 2.1 and, therefore, is consistent. Let ρ be the Mallows metric[2]. The Glivenko–Cantelli theorem and the strong law of large numbers imply that condition (i) of Theorem 2.1 is satisfied. The Lindeberg–Levy central limit theorem implies that T_n is asymptotically normally distributed. The cumulative normal distribution function is continuous, so condition (ii) holds. By using arguments similar to those used to prove the Lindeberg–Levy theorem, it can be shown that condition (iii) holds. ■

A theorem by Mammen (1992) gives necessary and sufficient conditions for the bootstrap to consistently estimate the distribution of a linear functional of F_0 when F_n is the EDF of the data. This theorem is important because the conditions are often easy to check, and many estimators and test statistics of interest in econometrics are asymptotically equivalent to linear functionals of some F_0. Hall (1990) and Gill (1989) give related theorems.

Theorem 2.2 (Mammen 1992). *Let $\{X_i : i = 1, \ldots, n\}$ be a random sample from a population. For a sequence of functions g_n and sequences of numbers t_n and σ_n, define $\bar{g}_n = n^{-1} \sum_{i=1}^n g_n(X_i)$ and $T_n = (\bar{g}_n - t_n)/\sigma_n$. For the bootstrap sample $\{X_i^* : i = 1, \ldots, n\}$, define $\bar{g}_n^* = n^{-1} \sum_{i=1}^n g_n(X_i^*)$ and $T_n^* = (\bar{g}_n^* - \bar{g}_n)/\sigma_n$. Let $G_n(\tau) = P(T_n \leqslant \tau)$ and $G_n^*(\tau) = P^*(T_n^* \leqslant \tau)$, where P^* is the probability distribution induced by bootstrap sampling. Then $G_n^*(\cdot)$ consistently estimates G_n if and only if $T_n \xrightarrow{d} N(0,1)$.* ■

If $E[g_n(X)]$ and $\mathrm{Var}[g_n(X)]$ exist for each n, then the asymptotic normality condition of Theorem 2.2 holds with $t_n = E(\bar{g}_n)$ and $\sigma_n^2 = \mathrm{Var}(\bar{g}_n)$ or $\sigma_n^2 = n^{-2} \sum_{i=1}^n [g_n(X_i) - \bar{g}_n]^2$. Thus, consistency of the bootstrap estimator of the distribution of the centered, normalized sample average in Example 2.1 follows trivially from Theorem 2.2.

The bootstrap need not be consistent if the conditions of Theorem 2.1 are not satisfied and is inconsistent if the asymptotic normality condition of Theorem 2.2 is not satisfied. In particular, the bootstrap tends to be inconsistent if F_0 is a point of discontinuity of the asymptotic distribution function $G_\infty(\tau, \cdot)$ or a point of superefficiency. Section 2.2 describes resampling methods that can sometimes be used to overcome these difficulties.

The following examples illustrate conditions under which the bootstrap is inconsistent. The conditions that cause inconsistency in the examples are unusual in econometric practice. The bootstrap is consistent in most applications. Nonetheless, inconsistency sometimes occurs, and it is important to be aware of its causes. Donald

[2] The Mallows metric is defined by $\rho(P,Q)^2 = \inf\left\{E\|Y - X\|^2 : Y \sim P, X \sim Q\right\}$. The infimum is over all joint distributions of (Y, X) whose marginals are P and Q. Weak convergence of a sequence of distributions in the Mallows metric implies convergence of the corresponding sequences of first and second moments. See Bickel and Freedman (1981) for a detailed discussion of this metric.

and Paarsch (1996), Flinn and Heckman (1982), and Heckman, Smith and Clements (1997) describe econometric applications that have features similar to those of some of the examples, though the consistency of the bootstrap in these applications has not been investigated.

Example 2.2. *Heavy-tailed distributions*: Let F_0 be the standard Cauchy distribution function and $\{X_i\}$ be a random sample from this distribution. Set $T_n = \bar{X}$, the sample average. Then T_n has the standard Cauchy distribution. Let F_n be the EDF of the sample. A bootstrap analog of T_n is $T_n^* = \bar{X}^* - m_n$, where $\bar{X}^*$ is the average of a bootstrap sample that is drawn randomly with replacement from the data $\{X_i\}$ and m_n is a median or trimmed mean of the data. The asymptotic normality condition of Theorem 2.2 is not satisfied, and the bootstrap estimator of the distribution of T_n is inconsistent. Athreya (1987) and Hall (1990) provide further discussion of the behavior of the bootstrap with heavy-tailed distributions. ■

Example 2.3. *The distribution of the square of the sample average*: Let $\{X_i: i=1,\dots,n\}$ be a random sample from a distribution with mean μ and variance σ^2. Let $\bar{X}$ denote the sample average. Let F_n be the EDF of the sample. Set $T_n = n^{1/2}(\bar{X}^2 - \mu^2)$ if $\mu \neq 0$ and $T_n = n\bar{X}^2$ otherwise. T_n is asymptotically normally distributed if $\mu \neq 0$, but T_n/σ^2 is asymptotically chi-square distributed with one degree of freedom if $\mu = 0$. The bootstrap analog of T_n is $T_n^* = n^a[(\bar{X}^*)^2 - \bar{X}^2]$, where $a=1/2$ if $\mu \neq 0$ and $a=1$ otherwise. The bootstrap estimator of $G_n(\tau, F_0) = P(T_n \leqslant \tau)$ is $G_n(\tau, F_n) = P_n^*(T_n^* \leqslant \tau)$. If $\mu \neq 0$, then T_n is asymptotically equivalent to a normalized sample average that satisfies the asymptotic normality condition of Theorem 2.2. Therefore, $G_n(\cdot, F_n)$ consistently estimates $G_\infty(\cdot, F_0)$ if $\mu \neq 0$. If $\mu = 0$, then T_n is not a sample average even asymptotically, so Theorem 2.2 does not apply. Condition (iii) of Theorem 2.1 is not satisfied, however, if $\mu = 0$, and it can be shown that the bootstrap distribution function $G_n(\cdot, F_n)$ does not consistently estimate $G_\infty(\cdot, F_0)$ [Datta (1995)]. ■

The following example is due to Bickel and Freedman (1981):

Example 2.4. *Distribution of the maximum of a sample*: Let $\{X_i: i=1,\dots,n\}$ be a random sample from a distribution with absolutely continuous CDF F_0 and support $[0, \theta_0]$. Let $\theta_n = \max(X_1,\dots,X_n)$, and define $T_n = n(\theta_n - \theta_0)$. Let F_n be the EDF of the sample. The bootstrap analog of T_n is $T_n^* = n(\theta_n^* - \theta_n)$, where θ_n^* is the maximum of the bootstrap sample $\{X_i^*\}$ that is obtained by sampling $\{X_i\}$ randomly with replacement. The bootstrap does not consistently estimate $G_n(-\tau, F_0) = P_n(T_n \leqslant -\tau)$ $(\tau \geqslant 0)$. To see why, observe that $P_n^*(T_n^* = 0) = 1 - (1 - 1/n)^n \to 1 - e^{-1}$ as $n \to \infty$. It is easily shown, however, that the asymptotic distribution function of T_n is $G_\infty(-\tau, F_0) = 1 - \exp[-\tau f(\theta_0)]$, where $f(x) = \mathrm{d}F(x)/\mathrm{d}x$ is the probability density function of X. Therefore, $P(T_n = 0) \to 0$, and the bootstrap estimator of $G_n(\cdot, F_0)$ is inconsistent. ■

Example 2.5. *Parameter on a boundary of the parameter space*: The bootstrap does not consistently estimate the distribution of a parameter estimator when the true parameter point is on the boundary of the parameter space. To illustrate, consider estimation of the population mean μ subject to the constraint $\mu \geqslant 0$. Estimate μ by $m_n = \bar{X}I(\bar{X} > 0)$, where $\bar{X}$ is the average of the random sample $\{X_i: i = 1, \ldots, n\}$. Set $T_n = n^{1/2}(m_n - \mu)$. Let F_n be the EDF of the sample. The bootstrap analog of T_n is $T_n^* = n^{1/2}(m_n^* - m_n)$, where m_n^* is the estimator of μ that is obtained from a bootstrap sample. The bootstrap sample is obtained by sampling $\{X_i\}$ randomly with replacement. If $\mu > 0$ and $\text{Var}(X) < \infty$, then T_n is asymptotically equivalent to a normalized sample average and is asymptotically normally distributed. Therefore, it follows from Theorem 2.2 that the bootstrap consistently estimates the distribution of T_n. If $\mu = 0$, then the asymptotic distribution of T_n is censored normal, and it can be proved that the bootstrap distribution function $G_n(\cdot, F_n)$ does not estimate $G_n(\cdot, F_0)$ consistently [Andrews (2000)]. ■

The next section describes resampling methods that often are consistent when the bootstrap is not. They provide consistent estimators of $G_n(\cdot, F_0)$ in each of the foregoing examples.

2.2. Alternative resampling procedures

This section describes two resampling methods whose requirements for consistency are weaker than those of the bootstrap. Each is based on drawing subsamples of size $m < n$ from the original data. In one method, the subsamples are drawn randomly with replacement. In the other, the subsamples are drawn without replacement. These subsampling methods often estimate $G_n(\cdot, F_0)$ consistently even when the bootstrap does not. They are not perfect substitutes for the bootstrap, however, because they tend to be less accurate than the bootstrap when the bootstrap is consistent.

In the first subsampling method, which will be called *replacement subsampling,* a bootstrap sample is obtained by drawing $m < n$ observations from the estimation sample $\{X_i: i = 1, \ldots, n\}$. In other respects, it is identical to the standard bootstrap based on sampling F_n. Thus, the replacement subsampling estimator of $G_n(\cdot, F_0)$ is $G_m(\cdot, F_n)$. Swanepoel (1986) gives conditions under which the replacement bootstrap consistently estimates the distribution of T_n in Example 2.4 (the distribution of the maximum of a sample). Andrews (2000) gives conditions under which it consistently estimates the distribution of T_n in Example 2.5 (parameter on the boundary of the parameter space). Bickel et al. (1997) provide a detailed discussion of the consistency and rates of convergence of replacement bootstrap estimators. To obtain some intuition into why replacement subsampling works, let F_{mn} be the EDF of a sample of size m drawn from the empirical distribution of the estimation data. Observe that if $m \to \infty$, $n \to \infty$, and $m/n \to 0$, then the random sampling error of F_n as an estimator of F_0 is smaller than the random sampling error of F_{mn} as an estimator of F_n. This makes the subsampling method less sensitive than the bootstrap to the behavior of $G_n(\cdot, F)$ for

F's in a neighborhood of F_0 and, therefore, less sensitive to violations of continuity conditions such as condition (iii) of Theorem 2.1.

The method of subsampling without replacement will be called *non-replacement subsampling.* This method has been investigated in detail by Politis and Romano (1994) and Politis et al. (1999), who show that it consistently estimates the distribution of a statistic under very weak conditions. In particular, the conditions required for consistency of the non-replacement subsampling estimator are much weaker than those required for consistency of the bootstrap estimator. Politis et al. (1997) extend the subsampling method to heteroskedastic time series.

To describe the non-replacement subsampling method, let $t_n = t_n(X_1, \ldots, X_n)$ be an estimator of the population parameter θ, and set $T_n = \rho(n)(t_n - \theta)$, where the normalizing factor $\rho(n)$ is chosen so that $G_n(\tau, F_0) = P(T_n \leqslant \tau)$ converges to a nondegenerate limit $G_\infty(\tau, F_0)$ at continuity points of the latter. In Example 2.1 (estimating the distribution of the sample average), for instance, θ is the population mean, $t_n = \bar{X}$, and $\rho(n) = n^{1/2}$. Let $\{X_{i_j}: j = 1, \ldots, m\}$ be a subset of $m < n$ observations taken from the sample $\{X_i: i = 1, \ldots, n\}$. Define $N_{mn} = \binom{n}{m}$ to be the total number of subsets that can be formed. Let $t_{m,k}$ denote the estimator t_m evaluated at the kth subset. The non-replacement subsampling method estimates $G_n(\tau, F_0)$ by

$$G_{nm}(\tau) \equiv \frac{1}{N_{nm}} \sum_{k=1}^{N_{nm}} I[\rho(m)(t_{m,k} - t_n) \leqslant \tau]. \tag{2.1}$$

The intuition behind this method is as follows. Each subsample $\{X_{i_j}\}$ is a random sample of size m from the population distribution whose CDF is F_0. Therefore, $G_m(\cdot, F_0)$ is the exact sampling distribution of $\rho(m)(t_m - \theta)$, and

$$G_m(\tau, F_0) = E\{I[\rho(m)(t_m - \theta) \leqslant \tau]\}. \tag{2.2}$$

The quantity on the right-hand side of Equation (2.2) cannot be calculated in an application because F_0 and θ are unknown. Equation (2.1) is the estimator of the right-hand side of Equation (2.2) that is obtained by replacing the population expectation by the average over subsamples and θ by t_n. If n is large but m/n is small, then random fluctuations in t_n are small relative to those in t_m. Accordingly, the sampling distributions of $\rho(m)(t_m - t_n)$ and $\rho(m)(t_m - \theta)$ are close. Similarly, if N_{mn} is large, the average over subsamples is a good approximation to the population average. These ideas are formalized in the following theorem of Politis and Romano (1994).

Theorem 2.3. *Assume that $G_n(\tau, F_0) \to G_\infty(\tau, F_0)$ as $n \to \infty$ at each continuity point of the latter function. Also assume that $\rho(m)/\rho(n) \to 0$, $m \to \infty$, and $m/n \to 0$ as $n \to \infty$. Let τ be a continuity point of $G_\infty(\tau, F_0)$. Then: (i) $G_{nm}(\tau) \xrightarrow{p} G_\infty(\tau, F_0)$; (ii) if $G_\infty(\cdot, F_0)$ is continuous, then $\sup_\tau |G_{nm}(\tau) - G_\infty(\tau, F_0)| \xrightarrow{p} 0$; (iii) let $c_n(1-\alpha) = \inf\{\tau: G_{nm}(\tau) \geqslant 1-\alpha\}$ and $c(1-\alpha, F_0) = \inf\{\tau: G_\infty(\tau, F_0) \geqslant 1-\alpha\}$. If $G_\infty(\cdot, F_0)$ is continuous at $c(1-\alpha, F_0)$, then $P[\rho(n)(t_n - \theta) \leqslant c_n(1-\alpha)] \to 1-\alpha$,* *and the asymptotic coverage probability of the confidence interval* $[t_n - \rho(n)^{-1} c_n(1-\alpha), \infty)$, *is* $1 - \alpha$.

Essentially, this theorem states that if T_n has a well-behaved asymptotic distribution, then the non-replacement subsampling method consistently estimates this distribution. The non-replacement subsampling method also consistently estimates asymptotic critical values for T_n and asymptotic confidence intervals for t_n.

In practice, N_{nm} is likely to be very large, which makes G_{nm} hard to compute. This problem can be overcome by replacing the average over all N_{nm} subsamples with the average over a random sample of subsamples [Politis and Romano (1994)]. These can be obtained by sampling the data $\{X_i\colon i=1,\ldots,n\}$ randomly without replacement.

It is not difficult to show that the conditions of Theorem 2.3 are satisfied in all of the statistics considered in Examples 2.1, 2.2, 2.4, and 2.5. The conditions are also satisfied by the statistic considered in Example 2.3 if the normalization constant is known. Bertail et al. (1999) describe a subsampling method for estimating the normalization constant $\rho(n)$ when it is unknown and provide Monte Carlo evidence on the numerical performance of the non-replacement subsampling method with an estimated normalization constant. In each of the foregoing examples, the replacement subsampling method works because the subsamples are random samples of the true population distribution of X, rather than an estimator of the population distribution. Therefore, replacement subsampling, in contrast to the bootstrap, does not require assumptions such as condition (iii) of Theorem 2.1 that restrict the behavior of $G_n(\cdot, F)$ for F's in a neighborhood of F_0.

The non-replacement subsampling method enables the asymptotic distributions of statistics to be estimated consistently under very weak conditions. However, the standard bootstrap is typically more accurate than non-replacement subsampling when the former is consistent. Suppose that $G_n(\cdot, F_0)$ has an Edgeworth expansion through $O(n^{-1/2})$, as is the case with the distributions of most asymptotically normal statistics encountered in applied econometrics. Then, as will be discussed in Section 3, $|G_n(\tau, F_n) - G_n(\tau, F_0)|$, the error made by the bootstrap estimator of $G_n(\tau, F_0)$, is at most $O(n^{-1/2})$ almost surely. In contrast, the error made by the non-replacement subsampling estimator, $|G_{nm}(\tau) - G_n(\tau, F_0)|$, is no smaller than $O_p(n^{-1/3})$ [Politis and Romano (1994), Politis et al. (1999)][3]. Thus, the standard bootstrap estimator of $G_n(\tau, F_0)$ is more accurate than the non-replacement subsampling estimator in a setting that arises frequently in applications. Similar results can be obtained for statistics that are asymptotically chi-square distributed. Thus, the standard bootstrap is more attractive than the non-replacement subsampling method in most applications in econometrics. The subsampling method may be used, however, if characteristics of the sampled population or the statistic of interest cause the standard bootstrap estimator

[3] Hall and Jing (1996) show how certain types of asymptotic refinements can be obtained through non-replacement subsampling. The rate of convergence of resulting error is, however, slower than the rate achieved with the standard bootstrap.

to be inconsistent. Non-replacement subsampling may also be useful in situations where checking the consistency of the bootstrap is difficult. Examples of this include inference about the parameters of certain kinds of structural search models [Flinn and Heckman (1982)], auction models [Donald and Paarsch (1996)], and binary-response models that are estimated by Manski's (1975, 1985) maximum score method.

3. Asymptotic refinements

The previous section described conditions under which the bootstrap yields a consistent estimator of the distribution of a statistic. Roughly speaking, this means that the bootstrap gets the statistic's asymptotic distribution right, at least if the sample size is sufficiently large. As was discussed in Section 1, however, the bootstrap often does much more than get the asymptotic distribution right. In a large number of situations that are important in applied econometrics, it provides a higher-order asymptotic approximation to the distribution of a statistic. This section explains how the bootstrap can be used to obtain asymptotic refinements. Section 3.1 describes the use of the bootstrap to reduce the finite-sample bias of an estimator. Section 3.2 explains how the bootstrap obtains higher-order approximations to the distributions of statistics. The results of Section 3.2 are used in Sections 3.3 and 3.4 to show how the bootstrap obtains higher-order refinements to the rejection probabilities of tests and the coverage probabilities of confidence intervals. Sections 3.5–3.7 address additional issues associated with the use of the bootstrap to obtain asymptotic refinements. It is assumed throughout this section that the data are a simple random sample from some distribution. Methods for implementing the bootstrap and obtaining asymptotic refinements with time-series data are discussed in Section 4.1.

3.1. Bias reduction

This section explains how the bootstrap can be used to reduce the finite-sample bias of an estimator. The theoretical results are illustrated with a simple numerical example. To minimize the complexity of the discussion, it is assumed that the inferential problem is to obtain a point estimate of a scalar parameter θ that can be expressed as a smooth function of a vector of population moments. It is also assumed that θ can be estimated consistently by substituting sample moments in place of population moments in the smooth function. Many important econometric estimators, including maximum-likelihood and generalized-method-of-moments estimators, are either functions of sample moments or can be approximated by functions of sample moments with an approximation error that approaches zero very rapidly as the sample size increases. Thus, the theory outlined in this section applies to a wide variety of estimators that are important in applications.

To be specific, let X be a random vector, and set $\mu = E(X)$. Assume that the true value of θ is $\theta_0 = g(\mu)$, where g is a known, continuous function. Suppose that the data consist of a random sample $\{X_i: i=1,\dots,n\}$ of X. Define the vector $\bar{X} = n^{-1}\sum_{i=1}^{n} X_i$. Then θ is estimated consistently by

$$\theta_n = g(\bar{X}). \tag{3.1}$$

If θ_n has a finite mean, then $E(\theta_n) = E[g(\bar{X})]$. However, $E[g(\bar{X})] \neq g(\mu)$ in general unless g is a linear function. Therefore, $E(\theta_n) \neq \theta_0$, and θ_n is a biased estimator of θ. In particular, $E(\theta_n) \neq \theta_0$ if θ_n is any of a variety of familiar maximum likelihood or generalized method of moments estimators.

To see how the bootstrap can reduce the bias of θ_n, suppose that g is four times continuously differentiable in a neighborhood of μ and that the components of X have finite fourth absolute moments. Let G_1 denote the vector of first derivatives of g and G_2 denote the matrix of second derivatives. A Taylor series expansion of the right-hand side of Equation (3.1) about $\bar{X} = \mu$ gives

$$\theta_n - \theta_0 = G_1(\mu)'(\bar{X} - \mu) + \tfrac{1}{2}(\bar{X} - \mu)' G_2(\mu)(\bar{X} - \mu) + R_n, \tag{3.2}$$

where R_n is a remainder term that satisfies $E(R_n) = \mathrm{O}(n^{-2})$. Therefore, taking expectations on both sides of Equation (3.2) gives

$$E(\theta_n - \theta_0) = \tfrac{1}{2}E[(\bar{X} - \mu)' G_2(\mu)(\bar{X} - \mu)] + \mathrm{O}(n^{-2}). \tag{3.3}$$

The first term on the right-hand side of Equation (3.3) has size $\mathrm{O}(n^{-1})$. Therefore, through $\mathrm{O}(n^{-1})$ the bias of θ_n is

$$B_n = \tfrac{1}{2}E[(\bar{X} - \mu)' G_2(\mu)(\bar{X} - \mu)]. \tag{3.4}$$

Now consider the bootstrap. The bootstrap samples the empirical distribution of the data. Let $\{X_i^*: i = 1, \dots, n\}$ be a bootstrap sample that is obtained this way. Define $\bar{X}^* = n^{-1}\sum_{i=1}^{n} X_i^*$ to be the vector of bootstrap sample means. The bootstrap estimator of θ is $\theta_n^* = g(\bar{X}^*)$. Conditional on the data, the true mean of the distribution sampled by the bootstrap is $\bar{X}$. Therefore, $\bar{X}$ is the bootstrap analog of μ, and $\theta_n = g(\bar{X})$ is the bootstrap analog of θ_0. The bootstrap analog of Equation (3.2) is

$$\theta_n^* - \theta_n = G_1(\bar{X})'(\bar{X}^* - \bar{X}) + \tfrac{1}{2}(\bar{X}^* - \bar{X})' G_2(\bar{X})(\bar{X}^* - \bar{X}) + R_n^*, \tag{3.5}$$

where R_n^* is the bootstrap remainder term. Let E^* denote the expectation under bootstrap sampling, that is, the expectation relative to the empirical distribution of the estimation data. Let $B_n^* \equiv E^*(\theta_n^* - \theta_n)$ denote the bias of θ_n^* as an estimator of θ_n. Taking E^* expectations on both sides of Equation (3.5) shows that

$$B_n^* = \tfrac{1}{2}E^*[(\bar{X}^* - \bar{X})' G_2(\bar{X})(\bar{X}^* - \bar{X})] + \mathrm{O}(n^{-2}) \tag{3.6}$$

almost surely. Because the distribution sampled by the bootstrap is known, B_n^* can be computed with arbitrary accuracy by Monte Carlo simulation. Thus, B_n^* is a feasible estimator of the bias of θ_n. The details of the simulation procedure are described below.

By comparing Equations (3.4) and (3.6), it can be seen that the only differences between B_n and the leading term of B_n^* are that $\bar{X}$ replaces μ in B_n^* and the empirical expectation, E^*, replaces the population expectation, E. Moreover, $E(B_n^*) = B_n + \mathrm{O}(n^{-2})$. Therefore, through $\mathrm{O}(n^{-1})$, use of the bootstrap bias estimate B_n^* provides the same bias reduction that would be obtained if the infeasible population value B_n could be used. This is the source of the bootstrap's ability to reduce the bias of θ_n. The resulting bias-corrected estimator of θ is $\theta_n - B_n^*$. It satisfies $E(\theta_n - \theta_0 - B_n^*) = \mathrm{O}(n^{-2})$. Thus, the bias of the bias-corrected estimator is $\mathrm{O}(n^{-2})$, whereas the bias of the uncorrected estimator θ_n is $\mathrm{O}(n^{-1})$[4].

The Monte Carlo procedure for computing B_n^* is as follows:

Monte Carlo Procedure for Bootstrap Bias Estimation

B1: Use the estimation data to compute θ_n.

B2: Generate a bootstrap sample of size n by sampling the estimation data randomly with replacement. Compute $\theta_n^* = g(\bar{X}^*)$.

B3: Compute $E^*\theta_n^*$ by averaging the results of many repetitions of step B2. Set $B_n^* = E^*\theta_n^* - \theta_n$.

To implement this procedure it is necessary to choose the number of repetitions, m, of step B2. It usually suffices to choose m sufficiently large that the estimate of $E^*\theta_n^*$ does not change significantly if m is increased further. Andrews and Buchinsky (2000) discuss more formal methods for choosing the number of bootstrap replications[5].

The following simple numerical example illustrates the bootstrap's ability to reduce bias. Examples that are more realistic but also more complicated are presented in Horowitz (1998a).

Example 3.1. *[Horowitz (1998a)]*: Let $X \sim N(0,6)$ and $n = 10$. Let $g(\mu) = \exp(\mu)$. Then $\theta_0 = 1$, and $\theta_n = \exp(\bar{X})$. B_n and the bias of $\theta_n - B_n^*$ can be found through the following Monte Carlo procedure:

MC1. Generate an estimation data set of size n by sampling from the $N(0,6)$ distribution. Use this data set to compute θ_n.

MC2. Compute B_n^* by carrying out steps B1–B3. Form $\theta_n - B_n^*$.

MC3. Estimate $E(\theta_n - \theta_0)$ and $E(\theta_n - B_n^* - \theta_0)$ by averaging the results of many repetitions of steps MC1–MC2. Estimate the mean square errors of θ_n and $\theta_n - B_n^*$ by averaging the realizations of $(\theta_n - \theta_0)^2$ and $(\theta_n - B_n^* - \theta_0)^2$.

[4] If $E(\theta_n)$ does not exist, then the "bias reduction" procedure described here centers a higher-order approximation to the distribution of $\theta_n - \theta_0$.

[5] It is not difficult to show that the bootstrap provides bias reduction even if $m = 1$. However, the bias-corrected estimator of θ may have a large variance if m is too small. The asymptotic distribution of the bias-corrected estimator is the same as that of the uncorrected estimator if m increases sufficiently rapidly as n increases. See Brown (1999) for further discussion.

The following are the results obtained with 1000 Monte Carlo replications and 100 repetitions of step B2 at each Monte Carlo replication:

	Bias	Mean-Square Error
θ_n	0.356	1.994
$\theta_n - B_n^*$	–0.063	1.246

In this example, the bootstrap reduces the magnitude of the bias of the estimator of θ by nearly a factor of 6. The mean-square estimation error is reduced by 38 percent. ■

3.2. The distributions of statistics

This section explains why the bootstrap provides an improved approximation to the finite-sample distribution of an asymptotically pivotal statistic. As before, the data are a random sample $\{X_i\colon i=1,\ldots,n\}$ from a probability distribution whose CDF is F_0. Let $T_n=T_n(X_1,\ldots,X_n)$ be a statistic. Let $G_n(\tau,F_0)=P(T_n\leqslant\tau)$ denote the exact, finite-sample CDF of T_n. As was discussed in Section 2, $G_n(\tau,F_0)$ cannot be calculated analytically unless T_n is pivotal. The objective of this section is to obtain an approximation to $G_n(\tau,F_0)$ that is applicable when T_n is not pivotal.

To obtain useful approximations to $G_n(\tau,F_0)$, it is necessary to make certain assumptions about the form of the function $T_n(X_1,\ldots,X_n)$. It is assumed in this section that T_n is a smooth function of sample moments of X or sample moments of functions of X (the smooth function model). Specifically $T_n = n^{1/2}[H(\bar{Z}_1, \ldots, \bar{Z}_J) - H(\mu_{Z_1}, \ldots, \mu_{Z_J})]$, where the scalar-valued function H is smooth in a sense that is defined precisely below, $\bar{Z}_j = n^{-1}\sum_{i=1}^{n} Z_j(X_i)$ for each $j=1,\ldots,J$ and some nonstochastic function Z_j, and $\mu_{Z_j} = E(Z_j)$. After centering and normalization, most estimators and test statistics used in applied econometrics are either smooth functions of sample moments or can be approximated by such functions with an approximation error that is asymptotically negligible[6]. The ordinary least-squares estimator of the slope coefficients in a linear mean-regression model and the t statistic for testing a hypothesis about a coefficient are exact functions of sample moments. Maximum-likelihood and generalized-method-of-moments estimators of the parameters of nonlinear models can be approximated with asymptotically negligible error by smooth functions of sample moments if the log-likelihood function or moment conditions have sufficiently many derivatives with respect to the unknown parameters.

[6] The meaning of asymptotic negligibility in this context may be stated precisely as follows. Let $\tilde{T}_n=\tilde{T}_n(X_1,\ldots,X_n)$ be a statistic, and let $T_n=n^{1/2}[H(\bar{Z}_1,\ldots,\bar{Z}_j)-H(\mu_{Z_1},\ldots,\mu_{Z_j})]$. Then the error made by approximating $\tilde{T}_n$ with T_n is asymptotically negligible if there is a constant $c>0$ such that $n^2P[n^2|\tilde{T}_n-T_n|>c]=\mathrm{O}(1)$ as $n\rightarrow\infty$.

Some important econometric estimators and test statistics do not satisfy the assumptions of the smooth function model. Quantile estimators, such as the least-absolute-deviations (LAD) estimator of the slope coefficients of a median-regression model do not satisfy the assumptions of the smooth function model because their objective functions are not sufficiently smooth. Nonparametric density and mean-regression estimators and semiparametric estimators that require kernel or other forms of smoothing also do not fit within the smooth function model. Bootstrap methods for such estimators are discussed in Section 4.3.

Now return to the problem of approximating $G_n(\tau, F_0)$. First-order asymptotic theory provides one approximation. To obtain this approximation, write $H(\bar{Z}_1, \ldots, \bar{Z}_J) = H(\bar{Z})$, where $\bar{Z} = (\bar{Z}_1, \ldots, \bar{Z}_J)'$. Define $\mu_Z = E(\bar{Z})$, $\partial H(z) = \partial H(z)/\partial z$, and $\Omega = E[(\bar{Z} - \mu_Z)(\bar{Z} - \mu_Z)']$ whenever these quantities exist. Assume that:

SFM: (i) $T_n = n^{1/2}[H(\bar{Z}) - H(\mu_Z)]$, *where* $H(z)$ *is 6 times continuously partially differentiable with respect to any mixture of components of* z *in a neighborhood of* μ_Z. (ii) $\partial H(\mu_Z) \neq 0$. (iii) *The expected value of the product of any 16 components of* Z *exists*[7].

Under assumption SFM, a Taylor series approximation gives

$$n^{1/2}[H(\bar{Z}) - H(\mu_Z)] = \partial H(\mu_Z)' n^{1/2}(\bar{Z} - \mu_Z) + o_p(1). \tag{3.7}$$

Application of the Lindeberg–Levy central limit theorem to the right hand side of Equation (3.7) shows that $n^{1/2}[H(\bar{Z}) - H(\mu_Z)] \xrightarrow{d} N(0, V)$, where $V = \partial H(\mu_Z)' \Omega \partial H(\mu_Z)$. Thus, the asymptotic CDF of T_n is $G_\infty(\tau, F_0) = \Phi(\tau/V^{1/2})$, where Φ is the standard normal CDF. This is just the usual result of the delta method. Moreover, it follows from the Berry–Esséen theorem that

$$\sup_\tau |G_n(\tau, F_0) - G_\infty(\tau, F_0)| = O(n^{-1/2}).$$

Thus, under assumption SFM of the smooth function model, first-order asymptotic approximations to the exact finite-sample distribution of T_n make an error of size $O(n^{-1/2})$[8].

[7] The proof that the bootstrap provides asymptotic refinements is based on an Edgeworth expansion of a sufficiently high-order Taylor-series approximation to T_n. Assumption SFM insures that H has derivatives and Z has moments of sufficiently high order to obtain the Taylor series and Edgeworth expansions that are used to obtain a bootstrap approximation to the distribution of T_n that has an error of size $O(n^{-2})$. SFM may not be the weakest condition needed to obtain this result. It certainly assumes the existence of more derivatives of H and moments of Z than needed to obtain less accurate approximations. For example, asymptotic normality of T_n can be proved if H has only one continuous derivative and Z has only two moments. See Hall (1992a, pp. 52–56; 238–259) for a statement of the regularity conditions needed to obtain various levels of asymptotic and bootstrap approximations.

[8] Some statistics that are important in econometrics have asymptotic chi-square distributions. Such statistics often satisfy the assumptions of the smooth function model but with $\partial H(\mu_Z) = 0$ and $\partial^2 H(z)/\partial z \partial z'|_{z=\mu_Z} \neq 0$. Versions of the results described here for asymptotically normal statistics are also available for asymptotic chi-square statistics. First-order asymptotic approximations to the finite-sample distributions of asymptotic chi-square statistics typically make errors of size $O(n^{-1})$. Chandra and Ghosh (1979) give a formal presentation of higher-order asymptotic theory for asymptotic chi-square statistics.

Now consider the bootstrap. The bootstrap approximation to the CDF of T_n is $G_n(\cdot, F_n)$. Under the smooth function model with assumption SFM, it follows from Theorem 3.2 that the bootstrap is consistent. Indeed, it is possible to prove the stronger result that $\sup_\tau |G_n(\tau, F_n) - G_\infty(\tau, F_0)| \to 0$ almost surely. This result insures that the bootstrap provides a good approximation to the asymptotic distribution of T_n if n is sufficiently large. It says nothing, however, about the accuracy of $G_n(\cdot, F_n)$ as an approximation to the exact finite-sample distribution function $G_n(\cdot, F_0)$. To investigate this question, it is necessary to develop higher-order asymptotic approximations to $G_n(\cdot, F_0)$ and $G_n(\cdot, F_n)$. The following theorem, which is proved in Hall (1992a), provides an essential result.

Theorem 3.1. *Let assumption SFM hold. Assume also that*

$$\lim \sup_{\|t\| \to \infty} |E[\exp(\mathrm{i}t'Z)]| < 1, \tag{3.8}$$

where $\mathrm{i} = \sqrt{-1}$. *Then*

$$G_n(\tau, F_0) = G_\infty(\tau, F_0) + \frac{1}{n^{1/2}} g_1(\tau, F_0) + \frac{1}{n} g_2(\tau, F_0) + \frac{1}{n^{3/2}} g_3(\tau, F_0) + \mathrm{O}(n^{-2}) \tag{3.9}$$

uniformly over τ *and*

$$G_n(\tau, F_n) = G_\infty(\tau, F_n) + \frac{1}{n^{1/2}} g_1(\tau, F_n) + \frac{1}{n} g_2(\tau, F_n) + \frac{1}{n^{3/2}} g_3(\tau, F_n) + \mathrm{O}(n^{-2}) \tag{3.10}$$

uniformly over τ *almost surely. Moreover,* g_1 *and* g_3 *are even, differentiable functions of their first arguments,* g_2 *is an odd, differentiable, function of its first argument, and* G_∞, g_1, g_2, *and* g_3 *are continuous functions of their second arguments relative to the supremum norm on the space of distribution functions.*

If T_n is asymptotically pivotal, then G_∞ is the standard normal distribution function. Otherwise, $G_\infty(\cdot, F_0)$ is the $N(0, V)$ distribution function, and $G_\infty(\cdot, F_n)$ is the $N(0, V_n)$ distribution function, where V_n is the quantity obtained from V by replacing population expectations and moments with expectations and moments relative to F_n.

Condition (3.8) is called the *Cramér condition.* It is satisfied if the random vector Z has a probability density with respect to Lebesgue measure[9].

[9] More generally, Equation (3.8) is satisfied if the distribution of Z has a non-degenerate absolutely continuous component in the sense of the Lebesgue decomposition. There are also circumstances in which Equation (3.8) is satisfied even when the distribution of Z does not have a non-degenerate absolutely continuous component. See Hall (1992a, pp. 66–67) for examples. In addition, Equation (3.8) can be modified to deal with econometric models that have a continuously distributed dependent variable but discrete covariates. See Hall (1992a, p. 266).

It is now possible to evaluate the accuracy of the bootstrap estimator $G_n(\tau, F_n)$ as an approximation to the exact, finite-sample CDF $G_n(\tau, F_0)$. It follows from Equations (3.9) and (3.10) that

$$\begin{aligned} G_n(\tau,F_n) - G_n(\tau,F_0) = {} & [G_\infty(\tau,F_n) - G_\infty(\tau,F_0)] + \frac{1}{n^{1/2}}[g_1(\tau,F_n) - g_1(\tau,F_0)] \\ & + \frac{1}{n}[g_2(\tau,F_n) - g_2(\tau,F_0)] + \mathrm{O}(n^{-3/2}) \end{aligned} \tag{3.11}$$

almost surely uniformly over τ. The leading term on the right-hand side of Equation (3.11) is $[G_\infty(\tau, F_n) - G_\infty(\tau, F_0)]$. The size of this term is $\mathrm{O}(n^{-1/2})$ almost surely uniformly over τ because $F_n - F_0 = \mathrm{O}(n^{-1/2})$ almost surely uniformly over the support of F_0. Thus, the bootstrap makes an error of size $\mathrm{O}(n^{-1/2})$ almost surely, which is the same as the size of the error made by first-order asymptotic approximations. In terms of rate of convergence to zero of the approximation error, the bootstrap has the same accuracy as first-order asymptotic approximations. In this sense, nothing is lost in terms of accuracy by using the bootstrap instead of first-order approximations, but nothing is gained either.

Now suppose that T_n is *asymptotically pivotal.* Then the asymptotic distribution of T_n is independent of F_0, and $G_\infty(\tau, F_n) = G_\infty(\tau, F_0)$ for all τ. Equations (3.9) and (3.10) now yield

$$\begin{aligned} G_n(\tau,F_n) - G_n(\tau,F_0) = {} & \frac{1}{n^{1/2}}[g_1(\tau,F_n) - g_1(\tau,F_0)] \\ & + \frac{1}{n}[g_2(\tau,F_n) - g_2(\tau,F_0)] + \mathrm{O}(n^{-3/2}) \end{aligned} \tag{3.12}$$

almost surely. The leading term on the right-hand side of Equation (3.12) is $n^{-1/2}[g_1(\tau, F_n) - g_1\ (\tau, F_0)]$. It follows from continuity of g_1 with respect to its second argument that this term has size $\mathrm{O}(n^{-1})$ almost surely uniformly over τ. Now the bootstrap makes an error of size $\mathrm{O}(n^{-1})$, which is smaller as $n \to \infty$ than the error made by first-order asymptotic approximations. Thus, the bootstrap is more accurate than first-order asymptotic theory for estimating the distribution of a smooth asymptotically pivotal statistic.

If T_n is asymptotically pivotal, then the accuracy of the bootstrap is even greater for estimating the symmetrical distribution function $P(|T_n| \leqslant \tau) = G_n(\tau, F_0) - G_n(-\tau, F_0)$. This quantity is important for obtaining the RP's of symmetrical tests and the coverage probabilities of symmetrical confidence intervals. Let Φ denote the standard normal distribution function. Then, it follows from Equation (3.9) and the symmetry of g_1, g_2, and g_3 in their first arguments that

$$\begin{aligned} G_n(\tau,F_0) - G_n(-\tau,F_0) &= [G_\infty(\tau,F_0) - G_\infty(-\tau,F_0)] + \frac{2}{n}g_2(\tau,F_0) + \mathrm{O}(n^{-2}) \\ &= 2\Phi(\tau) - 1 + \frac{2}{n}g_2(\tau,F_0) + \mathrm{O}(n^{-2}). \end{aligned} \tag{3.13}$$

Similarly, it follows from Equation (3.10) that

$$
\begin{aligned}
G_n(\tau, F_n) - G_n(-\tau, F_n) &= [G_\infty(\tau, F_n) - G_\infty(-\tau, F_n)] + \frac{2}{n} g_2(\tau, F_n) + \mathrm{O}(n^{-2}) \\
&= 2\Phi(\tau) - 1 + \frac{2}{n} g_2(\tau, F_n) + \mathrm{O}(n^{-2})
\end{aligned}
\tag{3.14}
$$

almost surely. The remainder terms in Equations (3.13) and (3.14) are $\mathrm{O}(n^{-2})$ and not $\mathrm{O}(n^{-3/2})$ because the $\mathrm{O}(n^{-3/2})$ term of an Edgeworth expansion, $n^{-3/2} g_3(\tau, F)$, is an even function that, like g_1, cancels out in the subtractions used to obtain Equations (3.13) and (3.14) from Equations (3.9) and (3.10). Now subtract Equation (3.13) from Equation (3.14) and use the fact that $F_n - F_0 = \mathrm{O}(n^{-1/2})$ almost surely to obtain

$$
\begin{aligned}
&[\, G_n(\tau, F_n) - G_n(-\tau, F_n)] - [G_n(\tau, F_0) - G_n(-\tau, F_0)] \\
&\quad = \frac{2}{n} [g_2(\tau, F_n) - g_2(\tau, F_0)] + \mathrm{O}(n^{-2}) \\
&\quad = \mathrm{O}(n^{-3/2})
\end{aligned}
\tag{3.15}
$$

almost surely if T_n is asymptotically pivotal. Thus, the error made by the bootstrap approximation to the symmetrical distribution function $P(|T_n| \leqslant \tau)$ is $\mathrm{O}(n^{-3/2})$ compared to the error of $\mathrm{O}(n^{-1})$ made by first-order asymptotic approximations.

In summary, when T_n is asymptotically pivotal, the error of the bootstrap approximation to a one-sided distribution function is

$$
G_n(\tau, F_n) - G_n(\tau, F_0) = \mathrm{O}(n^{-1}) \tag{3.16}
$$

almost surely uniformly over τ. The error in the bootstrap approximation to a symmetrical distribution function is

$$
[\, G_n(\tau, F_n) - G_n(-\tau, F_n)] - [G_n(\tau, F_0) - G_n(-\tau, F_0)] = \mathrm{O}(n^{-3/2}) \tag{3.17}
$$

almost surely uniformly over τ. In contrast, the errors made by first-order asymptotic approximations are $\mathrm{O}(n^{-1/2})$ and $\mathrm{O}(n^{-1})$, respectively, for one-sided and symmetrical distribution functions. Equations (3.16) and (3.17) provide the basis for the bootstrap's ability to reduce the finite-sample errors in the RP's of tests and the coverage probabilities of confidence intervals. Section 3.3 discusses the use of the bootstrap in hypothesis testing. Confidence intervals are discussed in Section 3.4.

3.3. Bootstrap critical values for hypothesis tests

This section shows how the bootstrap can be used to reduce the errors in the RP's of hypothesis tests relative to the errors made by first-order asymptotic approximations.

Let T_n be a statistic for testing a hypothesis H_0 about the sampled population. Assume that under H_0, T_n is asymptotically pivotal and satisfies assumptions SFM and Equation (3.8). Consider a symmetrical, two-tailed test of H_0. This test rejects H_0 at the α level if $|T_n| > z_{n,\,\alpha/2}$, where $z_{n,\alpha/2}$, the exact, finite-sample, α-level critical value, is the $1-\alpha/2$ quantile of the distribution of T_n[10]. The critical value solves the equation

$$G_n(z_{n,\,\alpha/2}, F_0) - G_n(-z_{n,\,\alpha/2}, F_0) = 1 - \alpha. \tag{3.18}$$

Unless T_n is exactly pivotal, however, Equation (3.18) cannot be solved in an application because F_0 is unknown. Therefore, the exact, finite-sample critical value cannot be obtained in an application if T_n is not pivotal.

First-order asymptotic approximations obtain a feasible version of Equation (3.18) by replacing G_n with G_∞. Thus, the asymptotic critical value, $z_{\infty,\,\alpha/2}$, solves

$$G_\infty(z_{\infty,\,\alpha/2}, F_0) - G_\infty(-z_{\infty,\,\alpha/2}, F_0) = 1 - \alpha. \tag{3.19}$$

Since G_∞ is the standard normal distribution function when T_n is asymptotically pivotal, $z_{\infty,\,\alpha/2}$ can be obtained from tables of standard normal quantiles. Combining Equations (3.13), (3.18), and (3.19) gives

$$[G_\infty(z_{n,\,\alpha/2}, F_0) - G_\infty(-z_{n,\,\alpha/2}, F_0)] - [G_\infty(z_{\infty,\,\alpha/2}, F_0) - G_\infty(-z_{\infty,\,\alpha/2}, F_0)] = \mathrm{O}(n^{-1}),$$

which implies that $z_{n,\,\alpha/2} - z_{\infty,\,\alpha/2} = \mathrm{O}(n^{-1})$. Thus, the asymptotic critical value approximates the exact finite sample critical value with an error whose size is $\mathrm{O}(n^{-1})$.

The bootstrap obtains a feasible version of Equation (3.18) by replacing F_0 with F_n. Thus, the bootstrap critical value, $z^*_{n,\,\alpha/2}$, solves

$$G_n(z^*_{n,\,\alpha/2}, F_n) - G_n(-z^*_{n,\,\alpha/2}, F_n) = 1 - \alpha. \tag{3.20}$$

Equation (3.20)[11] usually cannot be solved analytically, but $z^*_{n,\,\alpha/2}$ can be estimated with any desired accuracy by Monte Carlo simulation. To illustrate, suppose, as often

[10] Another form of two-tailed test is the equal-tailed test. An equal-tailed test rejects H_0 if $T_n > z_{n,\alpha/2}$ or $T_n < z_{n,(1-\alpha/2)}$, where $z_{n,(1-\alpha/2)}$ is the $\alpha/2$-quantile of the finite-sample distribution of T_n. If the distribution of T_n is symmetrical about 0, then equal-tailed and symmetrical tests are the same. Otherwise, they are different. Most test statistics used in econometrics have symmetrical asymptotic distributions, so the distinction between equal-tailed and symmetrical tests is not relevant when the RP is obtained from first-order asymptotic theory. Many test statistics have asymmetrical finite-sample distributions however. Higher-order approximations to these distributions, such as the approximation provided by the bootstrap, are also asymmetrical. Therefore, the distinction between equal-tailed and symmetrical tests is important in the analysis of asymptotic refinements. Note that "symmetrical" in a symmetrical test refers to the way in which the critical value is obtained, not to the finite-sample distribution of T_n, which is asymmetrical in general.

[11] The empirical distribution of the data is discrete, so Equation (3.20) may not have a solution if F_n is the EDF of the data. However, Hall (1992a, pp. 283–286) shows that there is a solution at a point α_n whose difference from α decreases exponentially fast as $n \to \infty$. The error introduced into the analysis by ignoring the difference between α_n and α is $\mathrm{o}(n^{-2})$ and, therefore, negligible for purposes of the discussion in this chapter.

happens in applications, that T_n is an asymptotically normal, Studentized estimator of a parameter θ whose value under H_0 is θ_0. That is,

$$T_n = \frac{n^{1/2}(\theta_n - \theta_0)}{s_n},$$

where θ_n is the estimator of $\theta, n^{1/2}(\theta_n - \theta_0) \xrightarrow{d} N(0, \sigma^2)$ under H_0 and s_n^2 is a consistent estimator of σ^2. Then the Monte Carlo procedure for evaluating $z^*_{n,\alpha/2}$ is as follows:

Monte Carlo Procedure for Computing the Bootstrap Critical Value

T1: Use the estimation data to compute θ_n.

T2: Generate a bootstrap sample of size n by sampling the distribution corresponding to F_n. For example, if F_n is the EDF of the data, then the bootstrap sample can be obtained by sampling the data randomly with replacement. If F_n is parametric so that $F_n(\cdot) = F(\cdot, \theta_n)$ for some function F, then the bootstrap sample can be generated by sampling the distribution whose CDF is $F(\cdot, \theta_n)$. Compute the estimators of θ and σ from the bootstrap sample. Call the results θ_n^* and s_n^*. The bootstrap version of T_n is $T_n^* = n^{1/2}(\theta_n^* - \theta_n)/s_n^*$.

T3: Use the results of many repetitions of T2 to compute the empirical distribution of $|T_n^*|$. Set $z^*_{n,\alpha/2}$ equal to the $1-\alpha$ quantile of this distribution.

The foregoing procedure does not specify the number of bootstrap replications that should be carried out in step T3. In practice, it often suffices to choose a value sufficiently large that further increases have no important effect on $z^*_{n,\alpha/2}$. Hall (1986a) and Andrews and Buchinsky (2000) describe the results of formal investigations of the problem of choosing the number of bootstrap replications. Repeatedly estimating θ in step T2 can be computationally burdensome if θ_n is an extremum estimator. Davidson and MacKinnon (1999a) and Andrews (1999) show that the computational burden can be reduced by replacing the extremum estimator with an estimator that is obtained by taking a small number of Newton or quasi-Newton steps from the θ_n value obtained in step T1.

To evaluate the accuracy of the bootstrap critical value $z^*_{n,\alpha/2}$ as an estimator of the exact finite-sample critical value $z_{n,\alpha/2}$, combine Equations (3.13) and (3.18) to obtain

$$2\Phi(z_{n,\alpha/2}) - 1 + \frac{2}{n} g_2(z_{n,\alpha/2}, F_0) = 1 - \alpha + \mathrm{O}(n^{-2}). \tag{3.21}$$

Similarly, combining Equations (3.14) and (3.20) yields

$$2\Phi(z^*_{n,\alpha/2}) - 1 + \frac{2}{n} g_2(z^*_{n,\alpha/2}, F_n) = 1 - \alpha + \mathrm{O}(n^{-2}), \tag{3.22}$$

almost surely. Equations (3.21) and (3.22) can be solved to yield Cornish–Fisher expansions for $z_{n,\alpha/2}$ and $z^*_{n,\alpha/2}$. The results are [Hall (1992a, p. 111)]

$$z_{n,\alpha/2} = z_{\infty,\alpha/2} - \frac{1}{n}\frac{g_2(z_{\infty,\alpha/2}, F_0)}{\phi(z_{\infty,\alpha/2})} + \mathrm{O}(n^{-2}), \tag{3.23}$$

where ϕ is the standard normal density function, and

$$z^*_{n,\alpha/2} = z_{\infty,\alpha/2} - \frac{1}{n}\frac{g_2(z_{\infty,\alpha/2}, F_n)}{\phi(z_{\infty,\alpha/2})} + \mathrm{O}(n^{-2}), \tag{3.24}$$

almost surely. It follows from Equations (3.23) and (3.24) that

$$z^*_{n,\alpha/2} = z_{n,\alpha/2} + \mathrm{O}(n^{-3/2}), \tag{3.25}$$

almost surely. Thus, the bootstrap critical value for a symmetrical, two-tailed test differs from the exact, finite-sample critical value by $\mathrm{O}(n^{-3/2})$ almost surely. The bootstrap critical value is more accurate than the asymptotic critical value, $z_{\infty,\alpha/2}$, whose error is $\mathrm{O}(n^{-1})$.

Now consider the rejection probability of the test based on T_n when H_0 is true. With the exact but infeasible α-level critical value, the RP is $P(|T_n| > z_{n,\alpha/2}) = \alpha$. With the asymptotic critical value, the RP is

$$\begin{aligned} P(|T_n| > z_{\infty,\alpha/2}) &= 1 - [G_n(z_{\infty,\alpha/2}, F_0) - G_n(-z_{\infty,\alpha/2}, F_0)] \\ &= \alpha + \mathrm{O}(n^{-1}), \end{aligned} \tag{3.26}$$

where the last line follows from setting $\tau = z_{\infty,\alpha/2}$ in Equation (3.13). Thus, with the asymptotic critical value, the true and nominal RP's differ by $\mathrm{O}(n^{-1})$.

Now consider the RP with the bootstrap critical value, $P(|T_n| \geqslant z^*_{n,\alpha/2})$. Because $z^*_{n,\alpha/2}$ is a random variable, $P(|T_n| \geqslant z^*_{n,\alpha/2}) \neq 1 - [G_n(z^*_{n,\alpha/2}, F_0) - G_n(-z^*_{n,\alpha/2}, F_0)]$. This fact complicates the calculation of the difference between the true and nominal RP's with the bootstrap critical value. The calculation is outlined in the Appendix of this chapter. The result is that

$$P(|T_n| > z^*_{n,\alpha/2}) = \alpha + \mathrm{O}(n^{-2}). \tag{3.27}$$

In other words, the nominal RP of a symmetrical, two-tailed test with a bootstrap critical value differs from the true RP by $\mathrm{O}(n^{-2})$ when the test statistic is asymptotically pivotal. In contrast, the difference between the nominal and true RP's is $\mathrm{O}(n^{-1})$ when the asymptotic critical value is used.

The bootstrap does not achieve the same accuracy for one-tailed tests. For such tests, the difference between the nominal and true RP's with a bootstrap critical value is usually $\mathrm{O}(n^{-1})$, whereas the difference with asymptotic critical values is $\mathrm{O}(n^{-1/2})$. See Hall (1992a, pp. 102–103) for details. There are, however, circumstances in which the difference between the nominal and true RP's with a bootstrap critical value is $O(n^{-3/2})$. Hall (1992a, pp. 178–179) shows that this is true for a one-sided t test of a hypothesis about a slope (but not intercept) coefficient in a homoskedastic, linear, mean-regression model. Davidson and MacKinnon (1999b) show that it is true whenever T_n is asymptotically independent of $g_2(z_{\infty,\alpha/2}, F_n)$. They further show that many familiar test statistics satisfy this condition.

Tests based on statistics that are asymptotically chi-square distributed behave like symmetrical, two-tailed tests. Therefore, the differences between their nominal and true RP's under H_0 are $O(n^{-1})$ with asymptotic critical values and $O(n^{-2})$ with bootstrap critical values.

Singh (1981), who considered a one-tailed test of a hypothesis about a population mean, apparently was the first to show that the bootstrap provides a higher-order asymptotic approximation to the distribution of an asymptotically pivotal statistic. Singh's test was based on the standardized sample mean. Early papers giving results on higher-order approximations for Studentized means and for more general hypotheses and test statistics include Babu and Singh (1983, 1984), Beran (1988) and Hall (1986b, 1988).

3.4. Confidence intervals

Let θ be a population parameter whose true but unknown value is θ_0. Let θ_n be a $n^{1/2}$-consistent, asymptotically normal estimator of θ, and let s_n be a consistent estimator of the standard deviation of the asymptotic distribution of $n^{1/2}(\theta_n - \theta_0)$. Then an asymptotic $1-\alpha$ confidence interval for θ_0 is $\theta_n - z_{\infty,\alpha/2} s_n/n^{1/2} \leqslant \theta_0 \leqslant \theta_n + z_{\infty,\alpha/2} s_n/n^{1/2}$. Define $T_n = n^{1/2}(\theta_n - \theta_0)/s_n$. Then the coverage probability of the asymptotic confidence interval is $P(|T_n| \leqslant z_{\infty,\alpha/2})$. It follows from Equation (3.26) that the difference between the true coverage probability of the interval and the nominal coverage probability, $1-\alpha$, is $O(n^{-1})$.

If T_n satisfies the assumptions of Theorem 3.1, then the difference between the nominal and true coverage probabilities of the confidence interval can be reduced by replacing the asymptotic critical value with the bootstrap critical value $z^*_{n,\alpha/2}$. With the bootstrap critical value, the confidence interval is $\theta_n - z^*_{n,\alpha/2} s_n/n^{1/2} \leqslant \theta_0 \leqslant \theta_n + z^*_{n,\alpha/2} s_n/n^{1/2}$. The coverage probability of this interval is $P(|T_n| \leqslant z^*_{n,\alpha/2})$. By Equation (3.27), $P(|T_n| \leqslant z^*_{n,\alpha/2}) = 1 - \alpha + O(n^{-2})$, so the true and nominal coverage probabilities differ by $O(n^{-2})$ when the bootstrap critical value is used, whereas they differ by $O(n^{-1})$ when the asymptotic critical value is used.

Analogous results can be obtained for one-sided and equal-tailed confidence intervals. With asymptotic critical values, the true and nominal coverage probabilities of these intervals differ by $O(n^{-1/2})$. With bootstrap critical values, the differences are $O(n^{-1})$. In special cases such as the slope coefficients of homoskedastic, linear, mean-regressions, the differences with bootstrap critical values are $O(n^{-3/2})$.

The bootstrap's ability to reduce the differences between the true and nominal coverage probabilities of a confidence interval is illustrated by the following example, which is an extension of Example 3.1.

Example 3.2. *[Horowitz (1998a)]*: This example uses Monte Carlo simulation to compare the true coverage probabilities of asymptotic and bootstrap nominal 95% confidence intervals for θ_0 in the model of Example 3.1. The Monte Carlo procedure is:

MC4: Generate an estimation data set of size $n=10$ by sampling from the $N(0,6)$ distribution. Use this data set to compute θ_n.

MC5: Compute $z^*_{n,\alpha/2}$ by carrying out steps T2–T3 of Section 3.3. Determine whether θ_0 is contained in the confidence intervals based on the asymptotic and bootstrap critical values.

MC6: Determine the empirical coverage probabilities of the asymptotic and bootstrap confidence intervals from the results of 1000 repetitions of steps MC4–MC5.

The empirical coverage probability of the asymptotic confidence interval was 0.886 in this experiment, whereas the empirical coverage probability of the bootstrap interval was 0.943. The asymptotic coverage probability is statistically significantly different from the nominal probability of 0.95 ($p<0.01$), whereas the bootstrap coverage probability is not ($p>0.10$). ■

3.5. The importance of asymptotically pivotal statistics

The arguments in Sections 3.2–3.4 show that the bootstrap provides higher-order asymptotic approximations to distributions, RP's of tests, and coverage probabilities of confidence intervals based on smooth, asymptotically pivotal statistics. These include test statistics whose asymptotic distributions are standard normal or chi-square and, thus, most statistics that are used for testing hypotheses about the parameters of econometric models. Models that satisfy the required smoothness conditions include linear and nonlinear mean-regression models, error-components mean-regression models for panel data, logit and probit models that have at least one continuously distributed explanatory variable, and tobit models. The smoothness conditions are also satisfied by parametric sample-selection models in which the selection equation is a logit or probit model with at least one continuously distributed explanatory variable. Asymptotically pivotal statistics based on median-regression models do not satisfy the smoothness conditions. Bootstrap methods for such statistics are discussed in Section 4.3. The ability of the bootstrap to provide asymptotic refinements for smooth, asymptotically pivotal statistics provides a powerful argument for using them in applications of the bootstrap.

The bootstrap may also be applied to statistics that are not asymptotically pivotal, but it does not provide higher-order approximations to their distributions. Estimators of the structural parameters of econometric models (e.g., slope and intercept parameters, including regression coefficients; standard errors, covariance matrix elements, and autoregressive coefficients) usually are not asymptotically pivotal. The asymptotic distributions of centered structural parameter estimators are often normal with means of zero but have variances that depend on the unknown population distribution of the data. The errors of bootstrap estimates of the distributions of statistics that are not asymptotically pivotal converge to zero at the same rate as the errors made by first-order asymptotic approximations [12].

Higher-order approximations to the distributions of statistics that are not asymptotically pivotal can be obtained through the use of bootstrap iteration [Beran (1987, 1988); Hall (1992a)] or bias-correction methods [Efron (1987)]. Bias correction methods are not applicable to symmetrical tests and confidence intervals. Bootstrap iteration is discussed in Section 4.4. Bootstrap iteration is highly computationally intensive, which makes it unattractive when an asymptotically pivotal statistic is available.

3.6. The parametric versus the nonparametric bootstrap

The size of the error in the bootstrap estimate of a RP or coverage probability is determined by the size of $F_n - F_0$. Thus, F_n should be the most efficient available estimator. If F_0 belongs to a known parametric family $F(\cdot, \theta)$, $F(\cdot, \theta_n)$ should be used to generate bootstrap samples, rather than the EDF. Although the bootstrap provides asymptotic refinements regardless of whether $F(\cdot, \theta_n)$ or the EDF is used, the results of Monte Carlo experiments have shown that the numerical accuracy of the bootstrap tends to be much higher with $F(\cdot, \theta_n)$ than with the EDF. If the objective is to test a hypothesis H_0 about θ, further gains in efficiency and performance can be obtained by imposing the constraints of H_0 when obtaining the estimate θ_n.

To illustrate, consider testing the hypothesis H_0: $\beta_1 = 0$ in the Box–Cox regression model

$$Y^{(\lambda)} = \beta_0 + \beta_1 X + U, \tag{3.28}$$

where $Y^{(\lambda)}$ is the Box and Cox (1964) transformation of Y, X is an observed, scalar explanatory variable, U is an unobserved random variable, and β_0 and β_1 are parameters. Suppose that $U \sim N(0, \sigma^2)$ [13]. Then bootstrap sampling can be carried out in the following ways:

(1) Sample (Y, X) pairs from the data randomly with replacement.

[12] Under mild regularity conditions, the constant that multiplies the rate of convergence of the error of the bootstrap estimate of the distribution function of a non-asymptotically-pivotal statistic is smaller than the constant that multiplies the rate of convergence of the error that is made by the normal approximation. This need not happen, however, with the errors in the RP's of tests and coverage probabilities of confidence intervals. See Beran (1982) and Liu and Singh (1987).

[13] Strictly speaking, U cannot be normally distributed unless $\lambda = 0$ or 1, but the error made by assuming normality is negligibly small if the right-hand side of the model has a negligibly small probability of being negative. Amemiya and Powell (1981) discuss ways to avoid assuming normality.

(2) Estimate λ, β_0, and β_1 in Equation (3.28) by maximum likelihood, and obtain residuals $\hat{U}$. Generate Y values from $Y = [\lambda_n(b_0 + b_1X + U^*) + 1]^{1/\lambda_n}$, where λ_n, b_0, and b_1 are the estimates of λ, β_0, and β_1; and U^* is sampled randomly with replacement from the $\hat{U}$.

(3) Same as method 2 except U^* is sampled randomly from the distribution $N(0, s_n^2)$, where s_n^2 is the maximum likelihood estimate of σ^2.

(4) Estimate λ, β_0, and σ^2 in Equation (3.28) by maximum likelihood subject to the constraint $\beta_1 = 0$. Then proceed as in method 2.

(5) Estimate λ, β_0, and σ^2 in Equation (3.28) by maximum likelihood subject to the constraint $\beta_1 = 0$. Then proceed as in method 3.

In methods 2–5, the values of X may be fixed in repeated samples or sampled independently of $\hat{U}$ from the empirical distribution of X.

Method 1 provides the least efficient estimator of F_n and typically has the poorest numerical accuracy. Method 5 has the greatest numerical accuracy. Method 3 will usually have greater numerical accuracy than method 2. If the distribution of U is not assumed to belong to a known parametric family, then methods 3 and 5 are not available, and method 4 will usually have greater numerical accuracy than methods 1–2. Of course, parametric maximum likelihood cannot be used to estimate β_0, β_1, and λ if the distribution of U is not specified parametrically.

If the objective is to obtain a confidence interval for β_1 rather than to test a hypothesis, methods 4 and 5 are not available. Method 3 will usually provide the greatest numerical accuracy if the distribution of U is assumed to belong to a known parametric family, and method 2 if not.

One reason for the relatively poor performance of method 1 is that it does not impose the condition $E(U|X = x) = 0$. This problem is discussed further in Section 5.2, where heteroskedastic regression models are considered.

3.7. Recentering

The bootstrap provides asymptotic refinements for asymptotically pivotal statistics because, under the assumptions of the smooth function model, $\sup_\tau |G_n(\tau, F_n) - G_n(\tau, F_0)|$ converges to zero as $n \to \infty$ more rapidly than $\sup_\tau |G_\infty(\tau, F_0) - G_n(\tau, F_0)|$. One important situation in which this does not necessarily happen is generalized method of moments (GMM) estimation of an overidentified parameter when F_n is the EDF of the sample.

To see why, let θ_0 be the true value of a parameter θ that is identified by the moment condition $Eh(X, \theta) = 0$. Assume that $\dim(h) > \dim(\theta)$. If, as is often the case in applications, the distribution of X is not assumed to belong to a known parametric family, the EDF of X is the most obvious candidate for F_n. The sample analog of $Eh(X, \theta)$ is then

$$E^*h(X, \theta) = \frac{1}{n}\sum_{i=1}^{n} h(X_i, \theta),$$

where E^* denotes the expectation relative to F_n. The sample analog of θ_0 is θ_n, the GMM estimator of θ. In general, $E^* h(X, \theta_n) \neq 0$ in an overidentified model, so bootstrap estimation based on the EDF of X implements a moment condition that does not hold in the population the bootstrap samples. As a result, the bootstrap estimator of the distribution of the statistic for testing the overidentifying restrictions is inconsistent [Brown et al. (1997)]. The bootstrap does consistently estimate the distributions of $n^{1/2}(\theta_n - \theta_0)$ [Hahn (1996)] and the t statistic for testing a hypothesis about a component of θ. However, it does not provide asymptotic refinements for the RP of the t test or the coverage probability of a confidence interval.

This problem can be solved by basing bootstrap estimation on the recentered moment condition $E^* h^*(X, \theta_n) = 0$, where

$$h^*(X, \theta) = h(X, \theta) - \frac{1}{n} \sum_{i=1}^{n} h(X_i, \theta_n). \tag{3.29}$$

Hall and Horowitz (1996) show that the bootstrap with recentering provides asymptotic refinements for the RP's of t tests of hypotheses about components of θ and the test of overidentifying restrictions. The bootstrap with recentering also provides asymptotic refinements for confidence intervals. Intuitively, the recentering procedure works by replacing the misspecified moment condition $E^* h(X, \theta) = 0$ with the condition $E^* h^*(X, \theta) = 0$, which does hold in the population that the bootstrap samples.

Freedman (1981) recognized the need for recentering residuals in regression models without intercepts. See also Efron (1979).

Brown et al. (1997) propose an alternative approach to recentering. Instead of replacing h with h^* for bootstrap estimation, they replace the empirical distribution of X with an empirical likelihood estimator that is constructed so that $E^* h(X, \theta_n) = 0$[14]. The empirical likelihood estimator assigns a probability mass π_{ni} to observation X_i $(i = 1, \ldots, n)$. The π_{ni}'s are determined by solving the problem

$$\underset{\pi_{n1}, \ldots, \pi_{nn}}{\text{maximize}} \sum_{i=1}^{n} \log \pi_{ni}$$

$$\text{subject to} \quad \sum_{i=1}^{n} \pi_{ni} h(X_i, \theta_n) = 0, \qquad \sum_{i=1}^{n} \pi_{ni} = 1, \quad \pi_{ni} \geqslant 0.$$

In general, the solution to this problem yields $\pi_{ni} \neq n^{-1}$, so the empirical likelihood estimator of the distribution of X is not the same as the empirical distribution. Brown et al. (1997) implement the bootstrap by sampling $\{X_i\}$ with probability weights π_{ni}

[14] The empirical-likelihood estimator is one of a larger class of estimators of F that are described by Brown et al. (1997) and that impose the restriction $E^* h(X, \theta_n) = 0$. All estimators in the class are asymptotically efficient.

instead of randomly with replacement. They argue that the bootstrap is more accurate with empirical-likelihood recentering than with recentering by Equation (3.29) because the empirical-likelihood estimator of the distribution of X is asymptotically efficient under the moment conditions $Eh(X, \theta) = 0$. With either method of recentering, however, the differences between the nominal and true RP's of symmetrical tests and between the nominal and true coverage probabilities of symmetrical confidence intervals are $O(n^{-2})$. Thus, the differences between the errors made with the two recentering methods are likely to be small with samples of the sizes typically encountered in applications.

Brown et al. (1997) develop the empirical-likelihood recentering method only for simple random samples. Kitamura (1997) has shown how to carry out empirical-likelihood estimation with dependent data. It is likely, therefore, that empirical-likelihood recentering can be extended to GMM estimation with dependent data. The recentering method based on Equation (3.29) requires no modification for use with dependent data [Hall and Horowitz (1996)]. Section 4.1 provides further discussion of the use of the bootstrap with dependent data.

4. Extensions

This section explains how the bootstrap can be used to obtain asymptotic refinements in certain situations where the assumptions of Section 3 are not satisfied. Section 4.1 treats dependent data. Section 4.2 treats kernel density and nonparametric mean-regression estimators. Section 4.3 shows how the bootstrap can be applied to certain non-smooth estimators. Section 4.4 describes how bootstrap iteration can be used to obtain asymptotic refinements without an asymptotically pivotal statistic. Section 4.5 discusses additional special problems that can arise in implementing the bootstrap. Section 4.6 discusses the properties of bootstrap critical values for testing a hypothesis that is false.

4.1. Dependent data

With dependent data, asymptotic refinements cannot be obtained by using independent bootstrap samples. Bootstrap sampling must be carried out in a way that suitably captures the dependence of the data-generation process. This section describes several methods for doing this. It also explains how the bootstrap can be used to obtain asymptotic refinements in GMM estimation with dependent data. At present, higher-order asymptotic approximations and asymptotic refinements are available only when the data-generation process is stationary and strongly geometrically mixing. Except when stated otherwise, it is assumed here that this requirement is satisfied. Non-stationary data-generation processes are discussed in Section 4.1.3.

4.1.1. Methods for bootstrap sampling with dependent data

Bootstrap sampling that captures the dependence of the data can be carried out relatively easily if there is a parametric model, such as an ARMA model, that reduces the data-generation process to a transformation of independent random variables. For example, suppose that the series $\{X_t\}$ is generated by the stationary, invertible, finite-order ARMA model

$$A(L,\alpha)X_t = B(L,\beta)\,U_t, \tag{4.1}$$

where A and B are known functions, L is the backshift operator, α and β are vectors of parameters, and $\{U_t\}$ is a sequence of independently and identically distributed (i.i.d.) random variables. Let α_n and β_n be $n^{1/2}$-consistent, asymptotically normal estimators of α and β, and let $\{\hat{U}_t\}$ be the centered residuals of the estimated model (4.1). Then a bootstrap sample $\{X_t^*\}$ can be generated as

$$A(L,\alpha_n)X_t^* = B(L,\beta_n)\,U_t^*,$$

where $\{U_t^*\}$ is a random sample from the empirical distribution of the residuals $\{\hat{U}_t\}$. If the distribution of U_t is assumed to belong to a known parametric family (e.g., the normal distribution), then $\{U_t^*\}$ can be generated by independent sampling from the estimated parametric distribution. Bose (1988) provides a rigorous discussion of the use of the bootstrap with autoregressions. Bose (1990) treats moving average models.

When there is no parametric model that reduces the data-generation process to independent sampling from some probability distribution, the bootstrap can be implemented by dividing the data into blocks and sampling the blocks randomly with replacement. The block bootstrap is important in GMM estimation with dependent data, because the moment conditions on which GMM estimation is based usually do not specify the dependence structure of the GMM residuals. The blocks may be non-overlapping [Carlstein (1986)] or overlapping [Hall (1985), Künsch (1989), Politis and Romano (1994)]. To describe these blocking methods more precisely, let the data consist of observations $\{X_i\colon i=1,\dots,n\}$. With non-overlapping blocks of length l, block 1 is observations $\{X_j\colon j=1,\dots,l\}$, block 2 is observations $\{X_{l+j}\colon j=1,\dots,l\}$, and so forth. With overlapping blocks of length l, block 1 is observations $\{X_j\colon j=1,\dots,l\}$, block 2 is observations $\{X_{j+1}\colon j=1,\dots,l\}$, and so forth. The bootstrap sample is obtained by sampling blocks randomly with replacement and laying them end-to-end in the order sampled. It is also possible to use overlapping blocks with lengths that are sampled randomly from the geometric distribution [Politis and Romano (1994)]. The block bootstrap with random block lengths is also called the *stationary bootstrap* because the resulting bootstrap data series is stationary, whereas it is not with overlapping or non-overlapping blocks of fixed (non-random) lengths.

Regardless of the blocking method that is used, the block length (or average block length in the stationary bootstrap) must increase with increasing sample size n to make bootstrap estimators of moments and distribution functions consistent. The asymptotically optimal block length is defined as the one that minimizes the asymptotic mean-square error of the block bootstrap estimator. The asymptotically optimal block length and its rate of increase with increasing n depend on what is being estimated.

Hall et al. (1995) showed that with either overlapping or non-overlapping blocks with non-random lengths, the asymptotically optimal block-length is $l \sim n^r$, where $r = 1/3$ for estimating bias or variance, $r = 1/4$ for estimating a one-sided distribution function (e.g., $P(T_n \leqslant \tau)$), and $r = 1/5$ for estimating a symmetrical distribution function (e.g., $P(|T_n| \leqslant \tau)$). Hall et al. (1995) also show that overlapping blocks provide somewhat higher estimation efficiency than non-overlapping ones. The efficiency difference is likely to be very small in applications, however. For estimating a two-sided distribution function, for example, the root-mean-square estimation error (RMSE) with either blocking method is $\mathrm{O}(n^{-6/5})$. The numerical difference between the RMSE's can be illustrated by considering the case of a normalized sample average. Let $T_n = (\bar{X} - \mu)/\sigma$, where $\bar{X}$ is the sample average of observations $\{X_i\}$, $\mu = E(\bar{X})$, and $\sigma^2 = \mathrm{Var}(\bar{X})$. Then the results of Hall et al. (1995) imply that for estimating $P(|T_n| \leqslant \tau)$, the reduction in asymptotic RMSE from using overlapping blocks instead of nonoverlapping ones is less than 10 percent.

Lahiri (1999) investigated the asymptotic efficiency of the stationary bootstrap. He showed that the asymptotic relative efficiency of the stationary bootstrap compared to the block bootstrap with non-random block lengths is always less than one and can be arbitrarily close to zero. More precisely, let $\mathrm{RMSE}_{\mathrm{SB}}$ and $\mathrm{RMSE}_{\mathrm{NR}}$, respectively, denote the asymptotic RMSE's of the stationary bootstrap and the block bootstrap with overlapping or non-overlapping blocks with non-random lengths. Then $\mathrm{RMSE}_{\mathrm{NR}}/\mathrm{RMSE}_{\mathrm{SB}} < 1$ always and can be arbitrarily close to zero. Thus, at least in terms of asymptotic RMSE, the stationary bootstrap is unattractive relative to the block bootstrap with fixed-length blocks.

Implementation of the block bootstrap in an application requires a method for choosing the block length with a finite sample. Hall et al. (1995) describe a subsampling method for doing this when the block lengths are non-random. The idea of the method is to use subsamples to create an empirical analog of the mean-square error of the bootstrap estimator of the quantity of interest. Let ψ denote this quantity (e.g., a two-sided distribution function). Let ψ_n be the bootstrap estimator of ψ that is obtained using a preliminary block-length estimate. Let $m < n$. Let $\psi_{m,i}(l')$ $(i = 1, \ldots, n-m)$ denote the bootstrap estimates of ψ that are computed using all the $n-m$ runs of length m in the data and block length l'. Let l_m be the value of l' that minimizes $\sum_i [\psi_{m,i}(l') - \psi_n]^2$. The estimator of the asymptotically optimal block length is $(n/m)^r l_m$, where $r = 1/3$ for estimating bias or variance, $r = 1/4$ for estimating a one-sided distribution function, and $r = 1/5$ for estimating a two-sided distribution function.

Kreiss (1992) and Bühlmann (1997) have proposed an alternative to blocking for use when the data-generation process can be represented as an infinite-order autoregression. In this method, called the *sieve bootstrap,* the infinite-order autoregression is replaced by an approximating autoregression with a finite-order that increases at a suitable rate as $n \to \infty$. The coefficients of the finite-order autoregression are estimated, and the bootstrap is implemented by sampling the centered residuals from the estimated finite-order model. Bühlmann (1997) gives conditions under which this procedure yields consistent estimators of variances and distribution functions. Bühlmann (1998) shows that the sieve bootstrap provides an asymptotic refinement for estimating the CDF of the t statistic for testing a one-sided hypothesis about the trend function in an AR(∞) process with a deterministic trend. Choi and Hall (2000) show that the error in the coverage probability of a one-sided confidence interval based on the sieve bootstrap for an AR(∞) process is $\mathrm{O}(n^{-1+\varepsilon})$ for any $\varepsilon > 0$, which is only slightly larger than the error of $\mathrm{O}(n^{-1})$ that is available when the data are a random sample.

If the data are generated by a Markov process, then the bootstrap can be implemented by sampling the process generated by a nonparametric estimate of the Markov transition density. This approach has been investigated by Rajarshi (1990), Datta and McCormick (1995), and Paparoditis and Politis (2000). Its ability to achieve asymptotic refinements for Studentized statistics is unknown.

4.1.2. Asymptotic refinements in GMM estimation with dependent data

This section discusses the use of the block bootstrap to obtain asymptotic refinements in GMM estimation with dependent data. Lahiri (1992) showed that the block bootstrap provides asymptotic refinements through $\mathrm{O}(n^{-1/2})$ for normalized sample moments and for a Studentized sample moment with m-dependent data. Hall and Horowitz (1996) showed that the block bootstrap provides asymptotic refinements through $\mathrm{O}(n^{-1})$ for symmetrical tests and confidence intervals based on GMM estimators. Their methods can also be used to show that the bootstrap provides refinements through $\mathrm{O}(n^{-1/2})$ for one-sided tests and confidence intervals. Hall and Horowitz (1996) do not assume that the data-generation process is m-dependent[15].

Regardless of whether overlapping or nonoverlapping blocks are used, block bootstrap sampling does not exactly replicate the dependence structure of the original data-generation process. For example, if nonoverlapping blocks are used, bootstrap observations that belong to the same block are deterministically related, whereas observations that belong to different blocks are independent. This dependence structure is unlikely to be present in the original data-generation process. As a result, the finite-sample covariance matrices of the asymptotic forms of parameter estimators obtained from the original sample and from the bootstrap sample are different. The practical consequence of this difference is that asymptotic refinements through $\mathrm{O}(n^{-1})$ cannot be obtained by applying the "usual" formulae for test statistics to the block-bootstrap sample. It is necessary to develop special formulae for the bootstrap versions of test statistics. These formulae contain factors that correct for the differences between the asymptotic covariances of the original-sample and bootstrap versions of

[15] The regularity conditions required to achieve asymptotic refinements in GMM estimation with dependent data include the existence of considerably more higher-order moments than are needed with i.i.d. data as well as a modified version of the Cramér condition that takes account of the dependence. See Hall and Horowitz (1996) for a precise statement of the conditions.

test statistics without distorting the higher-order terms of asymptotic expansions that produce refinements.

Lahiri (1992) derived the bootstrap version of a Studentized sample mean for m-dependent data. Hall and Horowitz (1996) derived formulae for the bootstrap versions of the GMM symmetrical, two-tailed t statistic and the statistic for testing overidentifying restrictions. As an illustration of the form of the bootstrap statistics, consider the GMM t statistic for testing a hypothesis about a component of a parameter θ that is identified by the moment condition $Eh(X, \theta) = 0$. Hall and Horowitz (1996) showed that the corrected formula for the bootstrap version of the GMM t statistic is

$$T_n^* = (S_n/S_b)\tilde{T}_n,$$

where $\tilde{T}_n$ is the "usual" GMM t statistic applied to the bootstrap sample, S_n is the "usual" GMM standard error of the estimate of the component of θ that is being tested, and S_b is the exact standard deviation of the asymptotic form of the bootstrap estimate of this component. S_n is computed from the original estimation sample, not the bootstrap sample. Hansen (1982) gives formulae for the usual GMM t statistic and standard error. S_b can be calculated because the process generating bootstrap data is known exactly. An analogous formula is available for the bootstrap version of the statistic for testing overidentifying restrictions but is much more complicated algebraically than the formula for the t statistic. See Hall and Horowitz (1996) for details.

At present, the block bootstrap is known to provide asymptotic refinements for symmetrical tests and confidence intervals based on GMM estimators only if the residuals $\{h(X_i, \theta_0): i = 1,2,\dots\}$ at the true parameter point, θ_0, are uncorrelated after finitely many lags. That is,

$$E[h(X_i, \theta_0)h(X_j, \theta_0)'] = 0 \quad \text{if} \quad |i-j| > M \tag{4.2}$$

for some $M < \infty$[16]. This restriction is not equivalent to m-dependence because it does not preclude correlations among higher powers of components of h that persist at arbitrarily large lags (e.g., stochastic volatility). Although the restriction is satisfied in many econometric applications [see, e.g., Hansen (1982), Hansen and Singleton (1982)], there are others in which relaxing it would be useful. The main problem in doing so is that without Equation (4.2), it is necessary to use a kernel-type estimator of the GMM covariance matrix [see, e.g., Newey and West (1987, 1994), Andrews (1991), Andrews and Monahan (1992)]. Kernel-type estimators are not functions of sample moments and converge at rates that are slower than $n^{-1/2}$. However, present

[16] Tests and confidence regions based on asymptotic chi-square statistics, including the test of overidentifying restrictions, are symmetrical. Therefore, restriction (4.2) also applies to them.

results on the existence of asymptotic expansions that achieve $O(n^{-1})$ accuracy with dependent data apply only to functions of sample moments that have $n^{-1/2}$ rates of convergence [Götze and Hipp (1983, 1994)]. It will be necessary to extend existing theory of asymptotic expansions with dependent data before Equation (4.2) can be relaxed for symmetrical tests and confidence intervals.

Condition (4.2) is not needed for one-sided tests and confidence intervals, where the bootstrap provides only $O(n^{-1/2})$ refinements. Götze and Künsch (1996) and Lahiri (1996) give conditions under which the moving-block-bootstrap approximation to the distribution of a statistic that is Studentized with a kernel-type variance estimator is accurate through $O_p(n^{-1/2})$. When the conditions are satisfied,

$$\sup_{\tau} |P(T_n \leqslant \tau) - P^*(T_n^* \leqslant \tau)| = o_p(n^{-1/2}), \tag{4.3}$$

where T_n^* is the bootstrap analog of the Studentized statistic T_n, and the moving block bootstrap is used to generate bootstrap samples. In Götze and Künsch (1996), T_n is the Studentized form of a smooth function of sample moments. In Lahiri (1996), T_n is a Studentized statistic for testing a hypothesis about a slope coefficient in a linear mean-regression model. Achieving the result (4.3) requires, among other things, use of a suitable kernel or weight function in the variance estimator. Götze and Künsch (1996) show that Equation (4.3) holds with a rectangular or quadratic kernel but not with a triangular one.

4.1.3. The bootstrap with non-stationary processes

The foregoing results assume that the data-generation process is stationary. Most research to date on using the bootstrap with non-stationary data has been concerned with establishing consistency of bootstrap estimators of distribution functions, not with obtaining asymptotic refinements. An exception is Lahiri (1992), who gives conditions under which the bootstrap estimator of the distribution of the normalized sample average of non-stationary data differs from the true distribution by $o(n^{-1/2})$ almost surely. Thus, under Lahiri's conditions, the bootstrap is more accurate than first-order asymptotic approximations. Lahiri's result requires *a priori* knowledge of the covariance function of the data and does not apply to Studentized sample averages. Moreover Lahiri assumes the existence of the covariance function, so his result does not apply to unit-root processes.

The consistency of the bootstrap estimator of the distribution of the slope coefficient or Studentized slope coefficient in a simple unit-root model has been investigated by Basawa et al. (1991a,b), Datta (1996), and Ferretti and Romo (1996). The model is

$$X_i = \beta X_{i-1} + U_i; \quad i = 1, 2, \ldots, n, \tag{4.4}$$

where $X_0 = 0$ and $\{U_i\}$ is an i.i.d. sequence with $E(U_i) = 0$ and $E(U_i^2) = \sigma^2 < \infty$. Let b_n denote the ordinary least-squares estimator of β in Equation (4.4):

$$b_n = \frac{\sum_{i=1}^{n} X_i X_{i-1}}{\sum_{i=1}^{n} X_{i-1}^2}. \tag{4.5}$$

Let β_0 denote the true but unknown value of β. Consider using the bootstrap to estimate the sampling distribution of $(b_n - \beta_0)$ or the t statistic for testing H$_0$: $\beta = \beta_0$. It turns out that when $\beta_0 = 1$ is possible, the consistency of the bootstrap estimator is much more sensitive to how the bootstrap sample is drawn than when it is known that $|\beta_0| < 1$.

Basawa et al. (1991a) investigate the consistency of a bootstrap estimator of the distribution of the t statistic in the special case that $U \sim N(0,1)$. In this case, the t statistic is

$$t_n = \left(\sum_{i=1}^{n} X_{i-1}^2 \right)^{1/2} (b_n - \beta_0). \tag{4.6}$$

In Basawa et al. (1991a), the bootstrap sample $\{X_i^*: i = 1, \ldots, n\}$ is generated recursively from the estimated model

$$X_i^* = b_n X_{i-1}^* + U_i^*, \tag{4.7}$$

where $X_0^* = 0$ and $\{U_i^*\}$ is an independent random sample from the $N(0,1)$ distribution. The bootstrap version of the t statistic is

$$t^* = \left(\sum_{i=1}^{n} (X_{i-1}^*)^2 \right)^{1/2} (b_n^* - b_n),$$

where b_n^* is obtained by replacing X_i with X_i^* in Equation (4.5). Basawa et al. (1991a) show that the bootstrap distribution function $P_n^*(t^* \leqslant \tau)$ does not consistently estimate the population distribution function $P_n(t \leqslant \tau)$. This result is not surprising. The asymptotic distribution of t is discontinuous at $\beta_0 = 1$. Therefore, condition (iii) of Theorem 2.1 is not satisfied if the set of data-generation processes under consideration includes ones with and without $\beta_0 = 1$.

This problem can be overcome by specifying that $\beta_0 = 1$, thereby removing the source of the discontinuity. Basawa et al. (1991b) investigate the consistency of the bootstrap estimator of the distribution of the statistic $Z_n \equiv n(b_n - 1)$ for testing the unit-root hypothesis H$_0$: $\beta_0 = 1$ in Equation (4.4). The bootstrap sample is generated by the recursion

$$X_i^* = X_{i-1}^* + U_i^*, \tag{4.8}$$

where $X_0^* = 0$ and $\{U_i^*\}$ is a random sample from the centered residuals of Equation (4.4) under H$_0$. The centered residuals are $\hat{U}_i = X_i - X_{i-1} - \bar{U}$, where $\bar{U} = n^{-1} \sum_{i=1}^{n} (X_i - X_{i-1})$. The bootstrap analog of Z_n is $Z_n^* = n(b_n^* - 1)$, where b_n^* is obtained by replacing X_i with X_i^* in Equation (4.5). Basawa et al. (1991b) show that if H$_0$ is true, then $|P_n^*(Z_n^* \leqslant z) - P_n(Z_n \leqslant z)| = o_p(1)$ uniformly over z.

The discontinuity problem can be overcome without the restriction $\beta_0 = 1$ by using bootstrap samples consisting of $m < n$ observations [Datta (1996)]. This approach has the advantage of yielding a confidence interval for β_0 that is valid for any $\beta_0 \in (-\infty, \infty)$. Consider model (4.4) with the additional assumption that $E|U_i|^{2+\delta} < \infty$ for some $\delta > 0$. Let b_n be the ordinary least-squares estimator of β, and define t_n as in Equation (4.6). Let $\hat{U}_i = X_i - b_n X_{i-1} - n^{-1} \sum_{i=1}^{n} (X_i - b_n X_{i-1})$ $(i = 1, \ldots, n)$ denote the centered residuals from the estimated model, and let $\{U_i^*: i = 1, \ldots, m\}$ be a random sample of $\{\hat{U}_i\}$ for some $m < n$. The bootstrap sample is generated by the recursion (4.7) but with $i = 1, \ldots, m$ instead of $i = 1, \ldots, n$. Let b_m^* denote the ordinary least-squares estimator of β that is obtained from the bootstrap sample. Define the bootstrap version of t_n by

$$t_m^* = \left(\sum_{i=1}^{m} (X_{i-1}^*)^2 \right)^{1/2} (b_m^* - b_n).$$

Datta (1996) proves that if $[m(\log \log n)^2]/n \to 0$ as $n \to \infty$, then $|P_m^*(t_m^* \leqslant \tau) - P_n(t_n \leqslant \tau)| = o(1)$ almost surely as $n \to \infty$ uniformly over z for any $\beta_0 \in (-\infty, \infty)$.

Ferretti and Romo (1996) consider a test of H_0: $\beta_0 = 1$ in Equation (4.4). Let b_n be the ordinary least-squares estimator of β, and let

$$\sigma_n^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - b_n X_{i-1})^2. \tag{4.9}$$

The test statistic is

$$\tilde{t}_n = \frac{1}{\sigma_n} \left(\sum_{i=1}^{n} X_{i-1}^2 \right)^{1/2} (b_n - 1). \tag{4.10}$$

The bootstrap sample is generated from the centered residuals of the estimated model by using the recursion (4.8). Let b_n^* denote the ordinary least-squares estimator of β that is obtained from the bootstrap sample. The bootstrap version of the test statistic, $\tilde{t}_n^*$, is obtained by replacing X_i and b_n with X_i^* and b_n^* in Equations (4.9) and (4.10). Ferretti and Romo (1996) show that $|P_n^*(\tilde{t}_n^* \leqslant \tau) - P_n(\tilde{t}_n \leqslant \tau)| = o(1)$, almost surely as $n \to \infty$. Ferretti and Romo (1996) also show how this result can be extended to the case in which $\{U_i\}$ in Equation (4.4) follows an AR(1) process.

The results of Monte Carlo experiments [Li and Maddala (1996, 1997)] suggest that the differences between the true and nominal RP's of tests of hypotheses about integrated or cointegrated data-generation processes are smaller with bootstrap-based critical values than with asymptotic ones. At present, however, there are no theoretical results on the ability of the bootstrap to provide asymptotic refinements for tests or confidence intervals when the data are integrated or cointegrated.

4.2. Kernel density and regression estimators

This section describes the use of the bootstrap to carry out inference about kernel nonparametric density and mean-regression estimators. These are not smooth functions of sample moments, even approximately, so the results of Section 3 do not apply to them. In particular, kernel density and mean-regression estimators converge more slowly than $n^{-1/2}$, and their distributions have unconventional asymptotic expansions that are not in powers of $n^{-1/2}$. Consequently, the sizes of the asymptotic refinements provided by the bootstrap are also not powers of $n^{-1/2}$. Sections 4.2.1–4.2.3 discuss bootstrap methods for nonparametric density estimation. Nonparametric mean regression is discussed in Section 4.2.4.

4.2.1. Nonparametric density estimation

Let f denote the probability density function (with respect to Lebesgue measure) of the scalar random variable X. The problem addressed in this section is inferring f from a random sample of X, $\{X_i\colon i=1,\dots,n\}$, without assuming that f belongs to a known, finite-dimensional family of functions. Point estimation of f can be carried out by the kernel method. The kernel estimator of $f(x)$ is

$$f_n(x) = \frac{1}{nh_n}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h_n}\right),$$

where K is a *kernel* function with properties that are discussed below and $\{h_n\colon n=1,2,\dots\}$ is a strictly positive sequence of bandwidths.

The properties of kernel density estimators are described by Silverman (1986), among others. To state the properties that are relevant here, let $r \geqslant 2$ be an even integer. Assume that f has r bounded, continuous derivatives in a neighborhood of x. Let K be a bounded function that is symmetrical about 0 and has support $[-1,1]$[17]. In addition, let K satisfy

$$\int_{-1}^{1} u^j K(u)\,\mathrm{d}u = \begin{cases} 1 \text{ if } j=0 \\ 0 \text{ if } 1 \leqslant j \leqslant r-1 \\ A_K \neq 0 \text{ if } j=r. \end{cases} \tag{4.11}$$

Define

$$B_K = \int_{-1}^{1} K(u)^2\,\mathrm{d}u.$$

Also define $b_n(x)=E[f_n(x)-f(x)]$ and $\sigma_n^2(x) = \mathrm{Var}[f_n(x)]$. Then

$$b_n(x) = h_n^r \frac{A_K}{r!} f^{(r)}(x) + \mathrm{o}(h_n^r)$$

[17] The results stated in this section do not require assuming that r is even or that K is a symmetrical function, but these assumptions simplify the exposition and are not highly restrictive in applications.

and

$$\sigma_n^2(x) = \frac{B_K}{nh_n} f(x) + \mathrm{o}[(nh_n^{-1})]. \tag{4.12}$$

Moreover, if nh_n^{2r+1} is bounded as $n \to \infty$, then

$$Z_n(x) \equiv \frac{f_n(x) - f(x) - b_n(x)}{\sigma_n(x)} = \frac{f_n(x) - E[f_n(x)]}{\sigma_n(x)} \xrightarrow{d} N(0, 1). \tag{4.13}$$

The fastest possible rate of convergence of $f_n(x)$ to $f(x)$ is achieved by setting $h_n \propto n^{-1/(2r+1)}$. When this happens, $f_n(x) - f(x) = \mathrm{O}_p[n^{-r/(2r+1)}]$, $b_n(x) \propto n^{-r/(2r+1)}$, and $\sigma_n(x) \propto n^{-r/(2r+1)}$.

A Studentized statistic that is asymptotically pivotal and can be used to test a hypothesis about $f(x)$ or form a confidence interval for $f(x)$ can be obtained from Equation (4.13) if suitable estimators of $\sigma_n^2(x)$ and $b_n(x)$ are available. The need for estimating an asymptotic variance is familiar. An estimator of $\sigma_n^2(x)$ can be formed by replacing $f(x)$ with $f_n(x)$ on the right-hand side of Equation (4.12). However, the asymptotic expansions required to obtain asymptotic refinements are simpler if $\sigma_n^2(x)$ is estimated by a sample analog of the exact, finite-sample variance of $f_n(x)$ instead of a sample analog of Equation (4.12), which is the variance of the asymptotic distribution of $f_n(x)$. A sample analog of the exact finite-sample variance of $f_n(x)$ is given by

$$s_n^2(x) = \frac{1}{(nh_n)^2} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right)^2 - \frac{f_n(x)^2}{n}.$$

If $h_n \to 0$ and $nh_n \to \infty$ as $n \to \infty$, then $(nh_n)[s_n^2(x) - \sigma_n^2(x)] = \mathrm{O}_p(1)$ as $n \to \infty$. Define the Studentized form of Z_n by

$$t_n = \frac{f_n(x) - E[f_n(x)]}{s_n(x)}. \tag{4.14}$$

Then t_n is the asymptotic t statistic for testing a hypothesis about $E[f_n(x)]$ or forming a confidence interval for $E[f_n(x)]$. The asymptotic distribution of t_n is $N(0, 1)$. However, unless the asymptotic bias $b_n(x)$ is negligibly small, t_n cannot be used to test a hypothesis about $f(x)$ or form a confidence interval for $f(x)$. Because $\sigma_n^{-1}(x) = \mathrm{O}[(nh_n)^{1/2}]$ and $s_n^{-1}(x) = \mathrm{O}_p[(nh_n)^{1/2}]$, $b_n(x)$ is negligibly small only if $(nh_n)^{1/2} b_n(x) = \mathrm{o}(1)$ as $n \to \infty$. The problem of asymptotic bias cannot be solved by replacing $E[f_n(x)]$ with $f(x)$ on the right-hand side of Equation (4.14) because the asymptotic distribution of the resulting version of t_n is not centered at 0 unless $b_n(x)$ is negligibly small. Section 4.2.2 discusses ways to deal with asymptotic bias.

4.2.2. Asymptotic bias and methods for controlling it

Asymptotic bias is a characteristic of nonparametric estimators that is not shared by estimators that are smooth functions of sample moments. As has just been explained, asymptotic bias may prevent t_n from being suitable for testing a hypothesis about $f(x)$ or constructing a confidence interval for $f(x)$. Asymptotic bias also affects the performance of the bootstrap. To see why, let $\{X_i^*: i=1,\dots,n\}$ be a bootstrap sample that is obtained by sampling the data $\{X_i\}$ randomly with replacement. Then the bootstrap estimator of f is

$$f_n^*(x) = \frac{1}{nh_n}\sum_{i=1}^{n} K\left(\frac{x-X_i^*}{h_n}\right). \tag{4.15}$$

The bootstrap analog of $s_n^2(x)$ is

$$s_n^{2*}(x) = \frac{1}{(nh_n)^2}\sum_{i=1}^{n} K\left(\frac{x-X_i^*}{h_n}\right)^2 - \frac{f_n^*(x)^2}{n}.$$

Define the bootstrap analog of t_n by

$$t_n^* = \frac{f_n^*(x)-f_n(x)}{s_n^*(x)}.$$

It is clear from Equation (4.15) that $E^*[f_n^*(x)-f_n(x)]=0$. Thus, $f_n^*(x)$ an unbiased estimator of $f_n(x)$ in a finite sample as well as asymptotically, whereas $f_n(x)$ is an asymptotically biased estimator of $f(x)$. It can be shown that the bootstrap distribution of t_n^* converges in probability to $N(0,1)$. Therefore, despite the unbiasedness of $f_n^*(x)$, t_n^* is a bootstrap t statistic for testing a hypothesis about $E[f_n(x)]$ or forming a confidence interval for $E[f_n(x)]$. It is not a bootstrap t statistic for testing a hypothesis about $f(x)$ or forming a confidence interval for $f(x)$ unless $b_n(x)$ is negligibly small.

There are two ways to overcome the difficulties posed by asymptotic bias so that t_n and t_n^* become statistics for testing hypotheses about $f(x)$ and forming confidence intervals for $f(x)$ instead of $E[f_n(x)]$. One is the method of explicit bias removal. It consists of forming an estimator of $b_n(x)$, say $\hat{b}_n(x)$, that can be subtracted from $f_n(x)$ to form the asymptotically unbiased estimator $f_n(x)-\hat{b}_n(x)$. The other method is undersmoothing. This consists of setting $h_n \propto n^{-\kappa}$ with $\kappa > 1/(2r+1)$. With undersmoothing, $(nh_n)^{1/2} b_n(x) = o_p(1)$ as $n \to \infty$, so that $b_n(x)$ is asymptotically negligible. Neither method is compatible with achieving the fastest rate of convergence of a point-estimator of $f(x)$. With undersmoothing, the rate of convergence of $f_n(x)$ is that of $\sigma_n(x)$. This is $n^{-(1-\kappa)/2}$, which is slower than $n^{-r/(2r+1)}$. Explicit bias removal with $h_n \propto n^{-1/(2r+1)}$ and rate of convergence $n^{-r/(2r+1)}$ for $f_n(x)$ requires $f(x)$ to have more than r derivatives. When $f(x)$ has the required number of derivatives, the fastest possible rate of convergence of $f_n(x)$ is $n^{-s/(2s+1)}$ for some $s > r$. This rate is achieved with $h_n \propto n^{-1/(2s+1)}$, but the resulting estimator of $f(x)$ is asymptotically biased. Thus, regardless of the method that is used to remove asymptotic bias, testing a hypothesis about $f(x)$ or forming a confidence interval requires using a bandwidth sequence that converges more rapidly than the one that maximizes the rate of convergence of a point estimator of $f(x)$. Nonparametric point estimation and nonparametric interval estimation or testing of hypotheses are different tasks that require different degrees of smoothing.

Hall (1992b) compares the errors in the coverage probabilities of bootstrap confidence intervals with undersmoothing and explicit bias removal. He shows that when the number of derivatives of $f(x)$ is held constant, undersmoothing achieves a smaller error in coverage probability than does explicit bias removal. This conclusion also applies to the rejection probabilities of hypothesis tests; the difference between true and nominal rejection probabilities can be made smaller with undersmoothing than with explicit bias removal. Thus, undersmoothing is the better method for handling asymptotic bias when the aim is to minimize differences between true and nominal rejection and coverage probabilities of bootstrap-based hypothesis tests and confidence intervals. Accordingly, undersmoothing is used for bias removal in the remainder of this section.

4.2.3. Asymptotic refinements

The argument showing that the bootstrap provides asymptotic refinements for tests of hypotheses and confidence intervals in nonparametric density estimation is similar to that made in Section 3 for the smooth function model. The main step is proving that the distributions of t_n and t_n^* have Edgeworth expansions that are identical up to a sufficiently small remainder. The result is stated in Theorem 4.1, which is proved in Hall (1992a, pp. 268–282).

Theorem 4.1. *Assume that f has r bounded, continuous derivatives in a neighborhood of x. Let $h_n \to 0$ and $(nh_n)/(\log n) \to \infty$ as $n \to \infty$. Let K be a bounded function that is symmetrical about 0, has support* $[-1, 1]$, *and satisfies Equation* (4.11) *for some $r \geqslant 2$. Also, assume that there is a partition of* $[-1, 1]$, $u_0 = -1 < u_1 < \ldots < u_m = 1$ *such that K' exists, is bounded, and is either strictly positive or strictly negative on each interval (u_j, u_{j+1}). Then there are even functions q_1 and q_3 and an odd function q_2 such that*

$$P(t_n \leqslant \tau) = \Phi(\tau) + \frac{1}{(nh_n)^{1/2}} q_1(\tau) + \frac{1}{nh_n} q_2(\tau) + \left(\frac{h_n}{n}\right)^{1/2} q_3(\tau) + \mathrm{O}[(nh_n)^{-3/2} + n^{-1}] \tag{4.16}$$

uniformly over τ. Moreover, there are even functions q_{n1} and q_{n3} and an odd function q_{n2} such that $q_{nj}(\tau) - q_j(\tau) \to 0$ as $n \to \infty$ uniformly over τ almost surely $(j = 1, \ldots, 3)$, *and*

$$P^*(t_n^* \leqslant \tau) = \Phi(\tau) + \frac{1}{(nh_n)^{1/2}} q_{n1}(\tau) + \frac{1}{nh_n} q_{n2}(\tau) + \left(\frac{h_n}{n}\right)^{1/2} q_{n3}(\tau) + \mathrm{O}[(nh_n)^{-3/2} + n^{-1}]$$

uniformly over τ almost surely.

Hall (1992a, pp. 211–216) gives explicit expressions for the functions q_j and q_{nj}.

To see the implications of Theorem 4.1, consider a symmetrical test of a hypothesis about $f(x)$. The results that will be obtained for this test also apply to symmetrical confidence intervals. Let the hypothesis be $\mathrm{H}_0{:}\,f(x)=f_0$. A symmetrical test rejects H_0 if $|f_n(x)-f_0|$ is large. Suppose that $nh_n^{r+1} \to 0$ as $n \to \infty$. This rate of convergence of h_n insures that the asymptotic bias of $f_n(x)$ has a negligibly small effect on the error made by the higher-order approximation to the distribution of t_n that is used to obtain asymptotic refinements[18]. It also makes the effects of asymptotic bias sufficiently small that t_n can be used to test H_0. Rejecting H_0 if $|f_n(x)-f_0|$ is large is then equivalent to rejecting H_0 if $|t_n|$ is large, thereby yielding a symmetrical t test of H_0.

Now suppose that the critical value of the symmetrical t test is obtained from the asymptotic distribution of t_n, which is $N(0,1)$. The asymptotic α-level critical value of the symmetrical t test is $z_{\alpha/2}$, the $1-\alpha/2$ quantile of the standard normal distribution. Theorem 4.1 shows that $P(|t_n| > z_{\alpha/2}) = \alpha + \mathrm{O}[(nh_n)^{-1}]$. In other words, when the asymptotic critical value is used, the difference between the true and nominal rejection probabilities of the symmetrical t test is $\mathrm{O}[(nh_n)^{-1}]$.

Now consider the symmetrical t test with a bootstrap critical value. The bootstrap α-level critical value, $z^*_{n,\alpha/2}$, satisfies $P^*(|t^*_n| \geqslant z^*_{n,\alpha/2}) = \alpha$. By Theorem 4.1,

$$P^*(|t^*_n| > \tau) - P(|t_n| > \tau) = \mathrm{o}[(nh_n)^{-1}] \tag{4.17}$$

almost surely uniformly over τ. It can also be shown that $P(|t_n| > z^*_{n,\alpha/2}) = \alpha + o[(nh_n)^{-1}]$. Thus, with the bootstrap critical value, the difference between the true and nominal rejection probabilities of the symmetrical t test is $\mathrm{o}[(nh_n)^{-1}]$. The bootstrap reduces the difference between the true and nominal rejection probabilities because it accounts for the effects of the $\mathrm{O}[(nh_n)^{-1}]$ term of the Edgeworth expansion of the distribution of t_n. First-order asymptotic approximations ignore this term. Thus, the bootstrap provides asymptotic refinements for hypothesis tests and confidence intervals based on a kernel nonparametric density estimator provided that the bandwidth h_n converges sufficiently rapidly to make the asymptotic bias of the density estimator negligibly small.

The conclusion that first-order asymptotic approximations make an error of size $\mathrm{O}[(nh_n)^{-1}]$ assumes that $nh_n^{r+1} \to 0$. If this condition is not satisfied, the error made by first-order approximations is dominated by the effect of asymptotic bias and is larger than $\mathrm{O}[(nh_n)^{-1}]$. This result is derived at the end of this section.

The bootstrap can also be used to obtain asymptotic refinements for one-sided and equal-tailed tests and confidence intervals. For one-sided tests and confidence intervals

[18] The asymptotic bias contributes a term of size $[(nh_n)^{1/2} b_n(x)]^2 = \mathrm{O}(nh_n^{2r+1})$ to the Edgeworth expansion of the distribution of $|t_n|$. Because t^*_n is unbiased, this term is not present in the expansion of the distribution of $|t^*_n|$. Therefore, the expansions of the distributions of $|t_n|$ and $|t^*_n|$ agree through $\mathrm{O}[(nh_n)^{-1}]$ only if $nh_n^{r+1} \to 0$ as $n \to \infty$.

with bootstrap critical values, the differences between the true and nominal rejection and coverage probabilities are $\mathrm{O}[(nh_n)^{-1} + (nh_n)^{1/2} h_n^r]$. These are minimized by setting $h_n \propto n^{-3/(2r+3)}$, in which case the errors are $\mathrm{O}[n^{-2r/(2r+3)}]$. For equal-tailed tests and confidence intervals with bootstrap critical values, the differences between the true and nominal rejection probabilities and coverage probabilities are $\mathrm{O}[(nh_n)^{-1} + nh_n^{2r+1} + h_n^r]$. These are minimized by setting $h_n \propto n^{-1/(r+1)}$, in which case the errors are $\mathrm{O}[n^{-r/(r+1)}]$. In contrast, the error made by first-order asymptotic approximations is $\mathrm{O}[(nh_n)^{-1/2}]$ in both the one-sided and equal-tailed cases. Hall (1992a, pp. 220–224) provides details and a discussion of certain exceptional cases in which smaller errors can be achieved. In contrast to the situation with the smooth function model, the orders of refinement achievable in nonparametric density estimation are different for one-sided and equal-tailed tests and confidence intervals.

4.2.3.1. The error made by first-order asymptotics when nh_n^{r+1} does not converge to 0. The effects of having $h_n \to 0$ too slowly are most easily seen by assuming that $\sigma_n(x)$ is known so that t_n is replaced by

$$Z_n = \frac{f_n(x) - f(x) - b_n(x)}{\sigma_n(x)}.$$

A symmetrical test of H_0 rejects if $|f_n(x) - f_0|/\sigma_n(x)$ is large. If H_0 is true, then

$$P\left(\frac{f_n(x) - f_0}{\sigma_n(x)} \leqslant \zeta\right) = P\left(Z_n \leqslant \zeta - \frac{b_n(x)}{\sigma_n(x)}\right)$$

for any ζ, and

$$P\left[\frac{|f_n(x) - f_0|}{\sigma_n(x)} \leqslant \zeta\right] = P\left[Z_n \leqslant \zeta - \frac{b_n(x)}{\sigma_n(x)}\right] - P\left[Z_n \leqslant -\zeta - \frac{b_n(x)}{\sigma_n(x)}\right]. \tag{4.18}$$

Each term on the right-hand side of Equation (4.18) has an asymptotic expansion of the form (4.16) except without the q_3 term and the $\mathrm{O}(n^{-1})$ remainder term, which arise from random sampling error in $s_n^2(x)$. Specifically,

$$\begin{aligned} P\left[\frac{|f_n(x) - f_0|}{\sigma_n(x)} \leqslant \zeta\right] &= \Phi\left[\zeta - \frac{b_n(x)}{\sigma_n(x)}\right] - \Phi\left[-\zeta - \frac{b_n(x)}{\sigma_n(x)}\right] \\ &\quad + \frac{1}{(nh_n)^{1/2}}\left\{p_1\left[\zeta - \frac{b_n(x)}{\sigma_n(x)}\right] - p_1\left[-\zeta - \frac{b_n(x)}{\sigma_n(x)}\right]\right\} \\ &\quad + \frac{1}{nh_n}\left\{p_2\left[\zeta - \frac{b_n(x)}{\sigma_n(x)}\right] - p_2\left[-\zeta - \frac{b_n(x)}{\sigma_n(x)}\right]\right\} + \mathrm{O}[(nh_n)^{-3/2}], \end{aligned} \tag{4.19}$$

where p_1 is an even function and p_2 is an odd function. Hall (1992a, p. 212) provides a proof and the details of p_1 and p_2. A Taylor series expansion of the right-hand side of Equation (4.19) combined with $b_n(x) = \mathrm{O}(h_n^r)$ and $\sigma_n(x) = \mathrm{O}[(nh_n)^{-1/2}]$ yields

$$P\left[\frac{|f_n(x) - f_0|}{\sigma_n(x)} \leqslant \zeta\right] = \Phi(\zeta) - \Phi(-\zeta) + \mathrm{O}[h_n^r + (nh_n)h_n^{2r} + (nh_n)^{-1}]. \tag{4.20}$$

The remainder term on the right-hand side of Equation (4.20) is dominated by h_n^r, which is the effect of asymptotic bias, unless $nh_n^{r+1} \to 0$. Thus, the error made by first-order asymptotic approximations exceeds $\mathrm{O}[(nh_n)^{-1}]$ unless $f_n(x)$ is sufficiently undersmoothed to make the asymptotic bias $b_n(x)$ negligible, which is equivalent to requiring $nh_n^{r+1} \to 0$ as $n \to \infty$.

4.2.4. Kernel nonparametric mean regression

In nonparametric mean-regression, the aim is to infer the mean of a random variable Y conditional on a covariate X without assuming that the conditional mean function belongs to a known finite-dimensional family of functions. Define $G(x) = E(Y|X = x)$ to be the conditional mean function. Let X be a scalar random variable whose distribution has a probability density function f. This section explains how the bootstrap can be used to obtain asymptotic refinements for tests of hypotheses about $G(x)$ and confidence intervals that are based on kernel estimation of G.

Let the data consist of a random sample, $\{Y_i, X_i\colon\ i = 1, \ldots, n\}$, of the joint distribution of (Y, X). The kernel nonparametric estimator of $G(x)$ is

$$G_n(x) = \frac{1}{nh_n f_n(x)} \sum_{i=1}^{n} Y_i K\left(\frac{x - X_i}{h_n}\right),$$

where

$$f_n(x) = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right),$$

K is a kernel function and $\{h_n\}$ a sequence of bandwidths. The properties of $G_n(x)$ are discussed by Härdle (1990). To state the ones that are relevant here, let $r \geqslant 2$ be an even integer. Assume that G and f each have r bounded, continuous derivatives in a neighborhood of x. Let K be a bounded function that is symmetrical about 0, has support $[-1, 1]$, and satisfies Equation (4.11). Define B_K and A_K as in Section 4.2.1.

Set $V(z) = \mathrm{Var}(Y|X = z)$, and assume that this quantity is finite and continuous in a neighborhood of $z = x$. Also define

$$b_n(x) = h_n^r \frac{A_K}{r!f(x)} \left\{ \frac{\partial^r}{\partial x^r} [G(x)f(x)] - f^{(r)}(x) \right\}$$

and

$$\sigma_n^2(x) = \frac{B_K}{nh_n} \frac{V(x)}{f(x)}. \tag{4.21}$$

If nh_n^{2r+1} is bounded as $n \to \infty$, then

$$Z_n(x) \equiv \frac{G_n(x) - G(x) - b_n(x)}{\sigma_n(x)} \xrightarrow{d} N(0, 1).$$

The fastest possible rate of convergence of $G_n(x)$ to $G(x)$ is achieved by setting $h_n \propto n^{-1/(2r+1)}$. When this happens, $G_n(x) - G(x) = \mathrm{O}_p[n^{-r/(2r+1)}]$, $b_n(x) \propto n^{-r/(2r+1)}$, and $\sigma_n(x) \propto n^{-r/(2r+1)}$.

The issues involved in converting Z_n into an asymptotically pivotal statistic that can be used to test a hypothesis about $G(x)$ or form a confidence interval for $G(x)$ are the same as in kernel density estimation. It is necessary to replace $\sigma_n(x)$ with a suitable estimator and to remove the asymptotic bias $b_n(x)$. As in kernel density estimation, asymptotic bias can be removed to sufficient order by undersmoothing. Undersmoothing for a symmetrical test or confidence interval consists of choosing h_n so that $nh_n^{r+1} \to 0$ as $n \to \infty$[19].

Now consider estimation of $\sigma_n^2(x)$. One possibility is to replace $f(x)$ with $f_n(x)$ and $V(x)$ with a consistent estimator on the right-hand side of Equation (4.21). The higher-order asymptotics of $G_n(x)$ are simpler, however, if $\sigma_n^2(x)$ is estimated by a sample analog of the exact finite-sample variance of the asymptotic form of $G_n(x) - G(x)$. With asymptotic bias removed by undersmoothing, the asymptotic form of $G_n(x) - G(x)$ is

$$G_n(x) - G(x) = \frac{1}{nh_n f(x)} \sum_{i=1}^{n} [Y_i - G(x)] K\left(\frac{x - X_i}{h_n}\right) + \mathrm{o}_p(1). \tag{4.22}$$

The variance of the first term on the right-hand side of Equation (4.22) is then estimated by the following sample analog, which will be used here to estimate $\sigma_n^2(x)$[20]:

$$s_n^2(x) = \frac{1}{[nh_n f_n(x)]^2} \sum_{i=1}^{n} [Y_i - G_n(x)]^2 K\left(\frac{x - X_i}{h_n}\right)^2.$$

[19] It is also possible to carry out explicit bias removal in kernel mean-regression. Härdle et al. (1995) compare the methods of explicit bias removal and undersmoothing for a one-sided confidence interval. They show that for a one-sided interval, there are versions of the bootstrap and explicit bias removal that give better coverage accuracy than the bootstrap with undersmoothing.

[20] Hall (1992a, p. 226) proposes an estimator of $\sigma_n^2(x)$ that is $n^{1/2}$-consistent when Y is homoskedastic (that is, $\mathrm{Var}(Y|X = x)$ is independent of x). The estimator used here is consistent (but not $n^{1/2}$-consistent) when Y has heteroskedasticity of unknown form.

Now define

$$t_n = \frac{G_n(x) - G(x)}{s_n(x)}.$$

With asymptotic bias removed through undersmoothing, t_n is asymptotically distributed as $N(0, 1)$ and is an asymptotically pivotal statistic that can be used to test a hypothesis about $G(x)$ and to form a confidence interval for $G(x)$. The bootstrap version of t_n is

$$t_n^* = \frac{G_n^*(x) - G_n(x)}{s_n^*(x)},$$

where $G_n^*(x)$ is obtained from $G_n(x)$ by replacing the sample $\{Y_i, X_i\}$ with the bootstrap sample $\{Y_i^*, X_i^*\}$, and $s_n^*(x)$ is obtained from $s_n(x)$ by replacing the sample with the bootstrap sample, $f_n(x)$ with $f_n^*(x)$, and $G_n(x)$ with $G_n^*(x)$[21].

The Edgeworth expansions of the distributions of t_n and t_n^* are similar in structure to those of the analogous statistic for kernel density estimators. The result for symmetrical tests and confidence intervals can be stated as follows. Let $E(Y^4|X = z)$ be finite and continuous for all z in a neighborhood of x. Let K satisfy the conditions of Theorem 4.1. Then there are functions q and q_n such that $q_n - q = \mathrm{o}(1)$ uniformly and almost surely as $n \to \infty$,

$$P(|t_n| \leqslant \tau) = 2\Phi(\tau) - 1 + \frac{1}{nh_n} q(\tau) + \mathrm{o}[(nh_n)^{-1}] \tag{4.23}$$

uniformly over τ, and

$$P^*(|t_n^*| \leqslant \tau) = 2\Phi(\tau) - 1 + \frac{1}{nh_n} q_n(\tau) + \mathrm{o}[(nh_n)^{-1}]$$

uniformly over τ almost surely. It follows that the bootstrap estimator of the distribution of $|t_n|$ is accurate through $\mathrm{O}[(nh_n)^{-1}]$, whereas first-order asymptotic approximations make an error of this size. Let $z_{n,\alpha/2}^*$ be the bootstrap α-level critical value for testing the hypothesis H_0: $G(x) = G_0$. Then $P^*(|t_n^*| > z_{n,\alpha/2}^*) = \alpha,$ and it can be shown that $P(|t_n| > z_{n,\alpha/2}^*) = \alpha + \mathrm{o}[(nh_n)^{-1}]$. Hall (1992a, Section 4.5) discusses the mathematical details. Thus, with the bootstrap critical value, the true and nominal rejection probabilities of a symmetrical t test of H_0 differ by $\mathrm{o}[(nh_n)^{-1}]$. In contrast, it follows from Equation (4.23) that the difference is $\mathrm{O}[(nh_n)^{-1}]$ if first-order asymptotic

[21] The discussion here assumes that the bootstrap sample is obtained by randomly sampling the empirical distribution of (Y, X). If $V(z)$ is a constant (that is, the model is homoskedastic), then bootstrap sampling can also be carried out by sampling centered regression residuals conditional on the observed values of X. See Hall (1992a, Section 4.5).

approximations are used to obtain the critical value. The same conclusions hold for the coverage probabilities of symmetrical confidence intervals for $G(x)$.

4.3. Non-smooth estimators

Some estimators are obtained by maximizing or minimizing a function that is discontinuous or whose first derivative is discontinuous. Two important examples are Manski's (1975, 1985) maximum-score (MS) estimator of the slope coefficients of a binary-response model and the least-absolute-deviations (LAD) estimator of the slope coefficients of a linear median-regression model. The objective function of the MS estimator and the first derivative of the objective function of the LAD estimator are step functions and, therefore, discontinuous. The LAD and MS estimators cannot be approximated by smooth functions of sample moments, so they do not satisfy the assumptions of the smooth function model. Moreover, the Taylor-series methods of asymptotic distribution theory do not apply to the LAD and MS estimators, which greatly complicates the analysis of their asymptotic distributional properties. As a consequence, little is known about the ability of the bootstrap to provide asymptotic refinements for hypothesis tests and confidence intervals based on these estimators. Indeed it is not known whether the bootstrap even provides a consistent approximation to the asymptotic distribution of the MS estimator.

This section explains how the LAD and MS estimators can be smoothed in a way that greatly simplifies the analysis of their asymptotic distributional properties. The bootstrap provides asymptotic refinements for hypothesis tests and confidence intervals based on the smoothed LAD and MS estimators. In addition, smoothing accelerates the rate of convergence of the MS estimator and simplifies even its first-order asymptotic distribution. Smoothing does not change the rate of convergence or first-order asymptotic distribution of the LAD estimator. The LAD estimator is treated in Section 4.3.1, and the MS estimator is treated in Section 4.3.2

4.3.1. The LAD estimator for a linear median-regression model

A linear median-regression model has the form

$$Y = X\beta + U, \tag{4.24}$$

where Y is an observed scalar, X is an observed $1\times q$ vector, β is a $q\times 1$ vector of constants, and U is an unobserved random variable that satisfies median$(U|X=x)=0$ almost surely. Let $\{Y_i, X_i\colon i=1,\ldots,n\}$ be a random sample from the joint distribution of (Y, X) in Equation (4.24). The LAD estimator of β, $\tilde{b}_n$, solves

$$\begin{aligned}\underset{b\,\in\,B}{\text{minimize}}\;\tilde{H}_n(b) &\equiv \frac{1}{n}\sum_{i=1}^{n}|Y_i - X_i b| \\ &= \frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i b)[2I(Y_i - X_i b > 0) - 1],\end{aligned} \tag{4.25}$$

where B is the parameter set and $I(\cdot)$ is the indicator function. Bassett and Koenker (1978) and Koenker and Bassett (1978) give conditions under which the LAD estimator is $n^{1/2}$-consistent and $n^{1/2}(\tilde{b}_n - \beta)$ is asymptotically normal.

$\tilde{H}_n(b)$ has cusps and, therefore, a discontinuous first derivative, at points b such that $Y_i = X_i b$ for some i. This non-smoothness causes the Edgeworth expansion of the LAD estimator to be non-standard and very complicated [De Angelis et al. (1993)]. The bootstrap is known to estimate the distribution of $n^{1/2}(\tilde{b}_n - \beta)$ consistently [De Angelis et al. (1993), Hahn (1995)], but it is not known whether the bootstrap provides asymptotic refinements for hypothesis tests and confidence intervals based on $\tilde{b}_n$[22].

Horowitz (1998b) suggests removing the cusps in $\tilde{H}_n$ by replacing the indicator function with a smooth function, thereby producing a modified objective function whose derivatives are continuous. The resulting smoothed LAD (SLAD) estimator is first-order asymptotically equivalent to the unsmoothed LAD estimator but has much simpler higher-order asymptotics. Specifically, let K be a bounded, differentiable function satisfying $K(v) = 0$ if $v \leqslant -1$ and $K(v) = 1$ if $v \geqslant 1$. Let $\{h_n\}$ be a sequence of bandwidths that converges to 0 as $n \to \infty$. The SLAD estimator solves

$$\underset{b \in B}{\text{minimize}}\ H_n(b) \equiv \frac{1}{n} \sum_{i=1}^{n} (Y_i - X_i b) \left[2K\left(\frac{Y_i - X_i b}{h_n} \right) - 1 \right]. \tag{4.26}$$

K is analogous to the integral of a kernel function for nonparametric density estimation. K is not a kernel function itself.

Let b_n be a solution to Equation (4.26). Horowitz (1998b) gives conditions under which $n^{1/2}(b_n - \tilde{b}_n) = o_p(1)$. Thus, the smoothed and unsmoothed LAD estimators are first-order asymptotically equivalent. It follows from this asymptotic equivalence and the asymptotic normality of LAD estimators that $n^{1/2}(b_n - \beta) \xrightarrow{d} N(0, V)$, where $V = D^{-1}E(X'X)D^{-1}$, $D = 2E[X'Xf(0|x)]$, and $f(\cdot|x)$ is the probability density function of U conditional on $X = x$.

A t statistic for testing a hypothesis about a component of β or forming a confidence interval can be constructed from consistent estimators of D and $E(X'X)$. D can be estimated consistently by $D_n(b_n)$, where

$$D_n(b) = \frac{2}{nh_n} \sum_{i=1}^{n} X_i' X_i K' \left(\frac{Y_i - X_i b}{h_n} \right). \tag{4.27}$$

$E(X'X)$ can be estimated consistently by the sample average of $X'X$. However, the asymptotic expansion of the distribution of the t statistic is simpler if $E(X'X)$ is

[22] Janas (1993) shows that a smoothed version of the bootstrap provides asymptotic refinements for a symmetrical t test of a hypothesis about a population median (no covariates).

estimated by the sample analog of the exact finite-sample variance of $\partial H_n(b)/\partial b$ at $b=\beta$. This estimator is $T_n(b_n)$, where

$$T_n(b) = \frac{1}{n}\sum_{i=1}^{n} X_i'X_i \left\{ \left[2K\left(\frac{Y_i - X_i b}{h_n}\right) - 1 \right] + 2\left(\frac{Y_i - X_i b}{h_n}\right) K'\left(\frac{Y_i - X_i b}{h_n}\right) \right\}^2 . \tag{4.28}$$

It is not difficult to show that V is estimated consistently by $V_n \equiv D_n(b_n)^{-1} T_n(b_n) D_n(b_n)^{-1}$. Now let b_{nj} and β_j, respectively, be the jth components of b_n and β ($j=1,\ldots,q$). Let V_{nj} be the (j,j) component of V_n. The t statistic for testing H_0: $\beta_j=\beta_{j0}$ is $t_n = n^{1/2}(b_{nj}-\beta_{j0})/V_{nj}^{1/2}$. If H_0 is true, then $t_n \xrightarrow{d} N(0,1)$, so t_n is asymptotically pivotal.

To obtain a bootstrap version of t_n, let $\{Y_i^*, X_i^*: i=1,\ldots,n\}$ be a bootstrap sample that is obtained by sampling the data $\{Y_i, X_i\}$ randomly with replacement. Let b_n^* be the estimator of β that is obtained by solving Equation (4.26) with $\{Y_i^*, X_i^*\}$ in place of $\{Y_i, X_i\}$. Let V_{nj}^* be the version of V_{nj} that is obtained by replacing b_n and $\{Y_i, X_i\}$, respectively, with b_n^* and $\{Y_i^*, X_i^*\}$ in Equations (4.27) and (4.28). Then the bootstrap analog of t_n is $t_n^* = n^{1/2}(b_{nj}^* - b_{nj})/(V_{nj}^*)^{1/2}$.

By using methods similar to those used with kernel density and mean-regression estimators, it can be shown that under regularity conditions, t_n and t_n^* have Edgeworth expansions that are identical almost surely through $\mathrm{O}[(nh_n)^{-1}]$. Horowitz (1998b) gives the details of the argument. In addition, reasoning similar to that used in Section 4.2.3 shows that the bootstrap provides asymptotic refinements for hypothesis tests and confidence intervals based on the SLAD estimator. For example, consider a symmetrical t test of H_0. Let $z_{n,\alpha/2}^*$ be the bootstrap α-level critical value for this test. That is, $z_{n,\alpha/2}^*$ satisfies $P^*(|t_n^*| > z_{n,\alpha/2}^*) = \alpha$. Then $P(|t_n| > z_{n,\alpha/2}^*) = \alpha + \mathrm{o}[(nh_n)^{-1}]$. In contrast, first-order asymptotic approximations make an error of size $\mathrm{O}[(nh_n)^{-1}]$. This is because first-order approximations ignore a term in the Edgeworth expansion of the distribution of $|t_n|$ whose size is $\mathrm{O}[(nh_n)^{-1}]$, whereas the bootstrap captures the effects of this term.

The conditions under which this result holds include: (1) for almost every x and every u in a neighborhood of 0, $f(u|x)$ is $r-1$ times continuously differentiable with respect to u; (2) K satisfies Equation (4.11) and has four bounded, Lipschitz continuous derivatives everywhere; and (3) $h_n \propto n^{-\kappa}$, where $2/(2r+1) < \kappa < 1/3$. Complete regularity conditions are given in Horowitz (1998b). Condition (3) implies that $r \geqslant 4$. Therefore, the size of the refinement obtained by the bootstrap is $\mathrm{O}(n^{-c})$, where $\frac{7}{9} < c < 1$.

The bootstrap also provides asymptotic refinements for one-sided tests and confidence intervals and for asymptotic chi-square tests of hypotheses about several components of β. In addition, it is possible to construct a smoothed version of Powell's (1984, 1986) censored LAD estimator and to show that the bootstrap provides asymptotic refinements for tests and confidence intervals based on the smoothed censored LAD estimator. Horowitz (1998b) provides details, a method for choosing h_n

in applications, and Monte Carlo evidence on the numerical performance of the t test with bootstrap critical values.

4.3.2. The maximum score estimator for a binary-response model

The most frequently used binary-response model has the form $Y = I(X\beta + U \geqslant 0)$, where X is an observed random vector, β is a conformable vector of constants, and U is an unobserved random variable. The parameter vector β is identified only up to scale, so a scale normalization is needed. Here, scale normalization will be accomplished by assuming that $|\beta_1| = 1$, where β_1 is the first component of β. Let $\tilde{\beta}$ and $\tilde{b}$ denote the vectors consisting of all components of β and b except the first. The maximum-score estimator of β, $b_n \equiv (b_{n1}, \tilde{b}_n')'$, solves

$$\underset{b \in B}{\text{maximize}}\ \tilde{H}_n(b) = \frac{1}{n}\sum_{i=1}^{n}(2Y_i - 1)I(X_i b \geqslant 0), \tag{4.29}$$

where $\{Y_i, X_i\colon i = 1, \ldots, n\}$ is a random sample from the joint distribution of (Y, X), and B is a compact parameter set in which the scale normalization holds.

Manski (1975, 1985) shows that if $\text{median}(U|X = x) = 0$ almost surely, the first component of X is continuously distributed with a non-zero coefficient, and certain other conditions are satisfied, then $(b_{n1}, \tilde{b}_n')' \rightarrow \beta$ almost surely. Because $b_{n1} = \pm 1$, b_{n1} converges to β_1 faster than any power of n. Cavanagh (1987) and Kim and Pollard (1990) show that $\tilde{b}_n$ converges in probability at the rate $n^{-1/3}$ and that $n^{1/3}(\tilde{b}_n - \tilde{\beta})$ has a complicated, non-normal asymptotic distribution. The MS estimator is important despite its slow rate of convergence and complicated limiting distribution because it is semiparametric (that is, it does not require the distribution of U to belong to a known, finite-dimensional family) and it permits the distribution of U to have arbitrary heteroskedasticity of unknown form provided that the centering assumption $\text{median}(U|X = x) = 0$ holds.

The asymptotic distribution of the MS estimator is too complex for use in testing hypotheses about β or constructing confidence intervals. Manski and Thompson (1986) suggested using the bootstrap to estimate the mean-square error of the MS estimator and presented Monte Carlo evidence suggesting that the bootstrap works well for this purpose. However, it is not known whether the bootstrap consistently estimates the asymptotic distribution of the MS estimator.

The MS estimator converges slowly and has a complicated limiting distribution because it is obtained by maximizing a step function. Horowitz (1992) proposed replacing the indicator function on the right-hand side of Equation (4.29) by a differentiable function. The resulting estimator is called the *smoothed maximum score* (SMS) estimator. It solves

$$\underset{b \in B}{\text{maximize}}\ H_n(b) = \frac{1}{n}\sum_{i=1}^{n}(2Y_i - 1)K\left(\frac{X_i b}{h_n}\right), \tag{4.30}$$

where K is a bounded, differentiable function satisfying $K(v)=0$ if $v \leqslant -1$ and $K(v)=1$ if $v \geqslant 1$, and $\{h_n\}$ is a sequence of bandwidths that converges to 0 as $n \to \infty$. As in SLAD estimation, K is analogous to the integral of a kernel function. Let $\tilde{\beta}$ again be the vector of all components of β but the first. Let $b_n \equiv (b_{n1}, \tilde{b}_n')'$ be the SMS estimator of $(\beta_1, \tilde{\beta}')'$. Horowitz (1992) gives conditions under which $(nh_n)^{1/2}(\tilde{b}_n - \tilde{\beta} - h_n^r \lambda) \xrightarrow{d} N(0, V)$, where $r \geqslant 2$ is an integer that is related to the number of times that the CDF of U and the density function of $X\beta$ are continuously differentiable, nh_n^{2r+1} is bounded as $n \to \infty$, λ is an asymptotic bias, and V is a covariance matrix. The rate of convergence of the SMS estimator of $\tilde{\beta}$ is at least $n^{-2/5}$ and can be arbitrarily close to $n^{-1/2}$ if the CDF of U and density function of $X\beta$ have sufficiently many derivatives. Thus, smoothing increases the rate of convergence of the MS estimator.

To obtain an asymptotically pivotal t statistic for testing a hypothesis about a component of $\tilde{\beta}$ or forming a confidence interval, it is necessary to remove the asymptotic bias of $\tilde{b}_n$ and construct a consistent estimator of V. Asymptotic bias can be removed by undersmoothing. For first-order asymptotic approximations, undersmoothing consists of choosing h_n so that $nh_n^{2r+1} \to 0$ as $n \to \infty$. However, for the reasons explained in the discussion of Equation (4.20), the stronger condition $nh_n^{r+1} \to 0$ is needed to obtain asymptotic refinements through $\mathrm{O}[(nh_n)^{-1}]$. V can be estimated consistently by $V_n = Q_n(b_n)^{-1} D_n(b_n) Q_n(b_n)^{-1}$, where for any $b \in B$

$$Q_n(b) = \frac{1}{nh_n^2} \sum_{i=1}^{n} (2Y_i - 1) \tilde{X}_i' \tilde{X}_i K'' \left(\frac{X_i b}{nh_n} \right), \tag{4.31}$$

$$D_n(b) = \frac{1}{nh_n} \sum_{i=1}^{n} \tilde{X}_i' \tilde{X}_i \left[K' \left(\frac{X_i b}{h_n} \right) \right]^2, \tag{4.32}$$

and $\tilde{X}$ consists of all components of X but the first.

Now let $\tilde{b}_{nj}$ and $\tilde{\beta}_j$, respectively, be the jth components of $\tilde{b}_n$ and $\tilde{\beta}$. Let V_{nj} be the (j,j) component of V_n. The t statistic for testing H_0: $\tilde{\beta}_j = \tilde{\beta}_{j0}$ is $t_n = (nh_n)^{1/2}(\tilde{b}_{nj} - \tilde{\beta}_{j0})/V_{nj}^{1/2}$. If H_0 is true, then $t_n \xrightarrow{d} N(0,1)$, so t_n is asymptotically pivotal.

To obtain a bootstrap version of t_n, let $\{Y_i^*, X_i^*: i = 1, \ldots, n\}$ be a bootstrap sample that is obtained by sampling the data $\{Y_i, X_i\}$ randomly with replacement. Let b_n^* be the estimator of β that is obtained by solving Equation (4.30) with $\{Y_i^*, X_i^*\}$ in place of $\{Y_i, X_i\}$. Let V_{nj}^* be the version of V_{nj} that is obtained by replacing b_n and $\{Y_i, X_i\}$, respectively, with b_n^* and $\{Y_i^*, X_i^*\}$ in Equations (4.31) and (4.32). Then the bootstrap analog of t_n is $t_n^* = (nh_n)^{1/2}(\tilde{b}_{nj}^* - \tilde{b}_{nj})/(V_{nj}^*)^{1/2}$.

By using methods similar to those used with kernel density and mean-regression estimators, it can be shown that t_n and t_n^* have Edgeworth expansions that are identical almost surely through $\mathrm{O}[(nh_n)^{-1}]$. See Horowitz (1998c) for the details of the argument. It follows that the bootstrap provides asymptotic refinements for hypothesis tests and confidence intervals based on the SMS estimator. For a symmetrical t test or confidence interval, the true and nominal rejection or coverage probabilities differ by $o[(nh_n)^{-1}]$ when bootstrap critical values are used, whereas they differ by $O[(nh_n)^{-1}]$ when first-order asymptotic critical values are used. First-order approximations ignore a term in the Edgeworth expansion of the distribution of $|t_n|$ whose size is $O[(nh_n)^{-1}]$, whereas the bootstrap captures the effects of this term.

The conditions under which this result holds include: (1) the CDF of U conditional on X and the density of $X\beta$ conditional on X have sufficiently many derivatives; (2) K satisfies Equation (4.11) for some $r \geqslant 8$; and (3) $h_n \propto n^{-\kappa}$, where $1/(r+1) < \kappa < 1/7$. Complete regularity conditions are given in Horowitz (1998c). Conditions (2) and (3) imply that the size of the refinement obtained by the bootstrap is $O(n^{-c})$, where $\frac{6}{7} < c < 1$. The bootstrap also provides asymptotic refinements for one-sided tests and confidence intervals and for asymptotic chi-square tests of hypotheses about several components of $\tilde{\beta}$. Horowitz (1998c) discusses methods for choosing h_n in applications and gives Monte Carlo evidence on the numerical performance of the t test with bootstrap critical values.

4.4. Bootstrap iteration

The discussion of asymptotic refinements in this chapter has emphasized the importance of applying the bootstrap to asymptotically pivotal statistics. This section explains how the bootstrap can be used to create an asymptotic pivot when one is not available. Asymptotic refinements can be obtained by applying the bootstrap to the bootstrap-generated asymptotic pivot. The computational procedure is called *bootstrap iteration* or *prepivoting* because it entails drawing bootstrap samples from bootstrap samples as well as using the bootstrap to create an asymptotically pivotal statistic. The discussion here concentrates on the use of prepivoting to test hypotheses [Beran (1988)]. Beran (1987) explains how to use prepivoting to form confidence regions. Hall (1986b) describes an alternative approach to bootstrap iteration.

Let T_n be a statistic for testing a hypothesis H_0 about a sampled population whose CDF is F_0. Assume that under H_0, T_n satisfies assumptions SFM and Equation (3.8) of the smooth function model. Define $F = F_0$ if H_0 is true, and define F to be the CDF of a distribution that satisfies H_0 otherwise. Let $G_n(\tau, F) \equiv P_F(T_n \leqslant \tau)$ denote the exact, finite-sample CDF of T_n under sampling from the population whose CDF is F. Suppose that H_0 is rejected if T_n is large. Then the exact α-level critical value of T_n, $z_{n\alpha}$, is the solution to $G_n(z_{n\alpha}, F) = 1 - \alpha$ under H_0. An exact α-level test based on T_n can be obtained by rejecting H_0 if $G_n(T_n, F) > 1 - \alpha$. Thus, if F were known, $g_n \equiv G_n(T_n, F)$ could be used as a statistic for testing H_0. Prepivoting is based on the idea of using g_n as a test statistic.

A test based on g_n cannot be implemented in an application unless T_n is pivotal because F and, therefore, g_n are unknown. A feasible test statistic can be obtained by replacing F with an estimator F_n that imposes the restrictions of H_0 and is $n^{1/2}$-consistent for F_0 if H_0 is true. Replacing F with F_n produces the bootstrap statistic $g_n^* = G_n(T_n, F_n)$. $G_n(\cdot, F_n)$ and, therefore, $G_n(T_n, F_n)$ can be estimated with arbitrary accuracy by carrying out a Monte Carlo simulation in which random samples are drawn from F_n. Given any τ, let $H_n(\tau, F_0) = P_{F_0}(g_n^* \leqslant \tau) = P_{F_0}[G_n(T_n, F_n) \leqslant \tau]$. An exact test based on g_n^* rejects H_0 at the α level if $H_n(g_n^*, F_0) > 1 - \alpha$. This test cannot be implemented because F_0 is unknown. If the bootstrap is consistent, however, the asymptotic distribution of g_n^* is uniform on $[0, 1]$. Therefore, H_0 is rejected at the asymptotic α level if $g_n^* > 1 - \alpha$. Now observe that g_n^* is asymptotically pivotal even if T_n is not; the asymptotic distribution of g_n^* is $U[0, 1]$ regardless of F_0. This suggests that asymptotic refinements can be obtained by carrying out a second stage of bootstrap sampling in which the bootstrap is used to estimate the finite-sample distribution of g_n^*.

The second stage of bootstrapping consists of drawing samples from each of the first-stage bootstrap samples that are used to compute g_n^*. Suppose that there are M first-stage samples. The mth such sample yields a bootstrap version of T_n, say T_{nm}, and an estimator F_{nm} of F_n that is consistent with H_0. F_{nm} can be sampled repeatedly to obtain $G_n(\cdot, F_{nm})$, the EDF of T_n under sampling from F_{nm}, and $g_{nm} \equiv G_n(T_{nm}, F_{nm})$. Now estimate $H_n(\cdot, F_0)$ by $H_n(\cdot, F_n)$, which is the EDF of g_{nm} $(m = 1, \ldots, M)$. The iterated bootstrap test rejects H_0 at the α level if $H_n(g_n^*, F_n) > 1 - \alpha$.

Beran (1988) shows that when prepivoting and bootstrap iteration are applied to a statistic T_n, the true and nominal probabilities of rejecting a correct null hypothesis differ by $\mathrm{o}(n^{-1/2})$ for a one-sided test and $\mathrm{o}(n^{-1})$ for a symmetrical test even if T_n is not asymptotically pivotal. By creating an asymptotic pivot in the first stage of bootstrapping, prepivoting and bootstrap iteration enable asymptotic refinements to be obtained for a non-asymptotically-pivotal T_n. The same conclusions apply to the coverage probabilities of confidence intervals. Beran (1988) presents the results of Monte Carlo experiments that illustrate the numerical performance of this procedure.

The computational procedure for carrying out prepivoting and bootstrap iteration is given by Beran (1988) and is as follows:

(1) Obtain T_n and F_n from the estimation data $\{X_i\colon i = 1, \ldots, n\}$, which are assumed to be a random sample of a possibly vector-valued random variable X.

(2) Let $\chi_1, \ldots, \chi_M$ be M bootstrap samples of size n that are drawn from the population whose distribution is F_n. Let F_{nm} denote the estimate of F_n that is obtained from χ_m. Let T_{nm} be the version of T_n that is obtained from χ_m. The EDF of $\{T_{nm}\colon m = 1, \ldots, M\}$ estimates $G_n(\cdot, F_n)$. Set $g_n^* = M^{-1} \sum_{m=1}^{M} I(T_{nm} \leqslant T_n)$.

(3) For each m, let $\chi_{m,1}, \ldots, \chi_{m,K}$ be K further bootstrap samples of size n, each drawn from the population whose CDF is F_{nm}. Let T_{nmk} be the version of T_n that is obtained from χ_{mk}. Set $G_n(T_{nm}, F_{nm}) = K^{-1} \sum_{k=1}^{K} I(T_{nmk} \leqslant T_{nm})$. Each of the $G_n(T_{nm}, F_{nm})$ $(m = 1, \ldots, n)$ is a second-stage estimate of g_n. Estimate $H_n(g_n^*, F_0)$ by $H_n(g_n^*, F_n) = M^{-1} \sum_{m=1}^{M} I[G_n(T_{nm}, F_{nm}) \leqslant g_n^*]$. Reject H_0 at the α level if $H_n(g_n^*, F_n) > 1 - \alpha$.

4.5. Special problems

The bootstrap provides asymptotic refinements because it amounts to a one-term Edgeworth expansion. The bootstrap cannot be expected to perform well when an Edgeworth expansion provides a poor approximation to the distribution of interest. An important case of this is instrumental-variables estimation with poorly correlated instruments and regressors. It is well known that first-order asymptotic approximations are especially poor in this situation [Hillier (1985), Nelson and Startz (1990a,b), Phillips (1983)]. The bootstrap does not offer a solution to this problem. With poorly correlated instruments and regressors, Edgeworth expansions of estimators and test statistics involve denominator terms that are close to zero. As a result, the higher-order terms of the expansions may dominate the lower-order ones for a given sample size, in which case the bootstrap may provide little improvement over first-order asymptotic approximations. Indeed, with small samples the numerical accuracy of the bootstrap may be even worse than that of first-order asymptotic approximations.

The bootstrap also does not perform well when the variance estimator used for Studentization has a high variance itself. This problem can be especially severe when the parameters being estimated or tested are variances or covariances of a distribution. This happens, for example, in estimation of covariance structures of economic processes [Abowd and Card (1987, 1989), Behrman et al. (1994), Griliches (1979), Hall and Mishkin (1982)]. In such cases Studentization is carried out with an estimator of the variance of an estimated variance. Imprecise estimation of a variance also affects the finite-sample performance of asymptotically efficient GMM estimators because the asymptotically optimal weight matrix is the inverse of the covariance matrix of the GMM residuals. The finite-sample mean-square error of the asymptotically efficient estimator can greatly exceed the mean-square error of an asymptotically inefficient estimator that is obtained with a non-stochastic weight matrix. Horowitz (1998a) shows that in the case of estimating covariance structures, this problem can be greatly mitigated by using a trimmed version of the covariance estimator that excludes "outlier" observations. See Horowitz (1998a) for details. Section 5.5 presents a numerical illustration of the effects of trimming.

4.6. The bootstrap when the null hypothesis is false

To understand the power of a test based on a bootstrap critical value, it is necessary to investigate the behavior of the bootstrap when the null hypothesis being tested, H_0, is false. Suppose that bootstrap samples are generated by a model that satisfies a false H_0 and, therefore, is misspecified relative to the true data-generation process. If H_0 is simple, meaning that it completely specifies the data-generation process, then the bootstrap amounts to Monte Carlo estimation of the exact finite-sample critical value for testing H_0 against the true data-generation process. Indeed, the bootstrap provides the exact critical value, rather than a Monte Carlo estimate, if $G(\cdot, F_n)$ can be calculated analytically. Tests of simple hypotheses are rarely encountered in econometrics, however.

In most applications, H_0 is composite. That is, it does not specify the value of a finite- or infinite-dimensional "nuisance" parameter ψ. In the remainder of this section, it is shown that a test of a composite hypothesis using a bootstrap-based critical value is a higher-order approximation to a certain exact test. The power of the test with a bootstrap critical value is a higher-order approximation to the power of the exact test.

Except in the case of a test based on a pivotal statistic, the exact finite-sample distribution of the test statistic depends on ψ. Therefore, except in the pivotal case, it is necessary to specify the value of ψ to obtain exact finite-sample critical values. The higher-order approximation to power provided by the bootstrap applies to a value of ψ that will be called the *pseudo-true value*. To define the pseudo-true value, let ψ_n be an estimator of ψ that is obtained under the incorrect assumption that H_0 is true. Under regularity conditions [see, e.g., Amemiya (1985), White (1982)], ψ_n converges in probability to a limit ψ^*, and $n^{1/2}(\psi_n - \psi^*) = O_p(1)$. ψ^* is the pseudo-true value of ψ.

Now let T_n be a statistic that is asymptotically pivotal under H_0. Suppose that its exact CDF with an arbitrary value of ψ is $G_n(\cdot, \psi)$, and that under H_0 its asymptotic CDF is $G_0(\cdot)$. Suppose that bootstrap sampling is carried out subject to the constraints of H_0. Then the bootstrap generates samples from a model whose parameter value is ψ_n, so the exact distribution of the bootstrap version of T_n is $G_n(\cdot, \psi_n)$. Under H_0 and subject to regularity conditions, $G_n(\cdot, \psi_n)$ has an asymptotic expansion of the form

$$G_n(z, \psi_n) = G_0(z) + n^{-j/2} g_j(z, \psi^*) + o_p(n^{-j/2}) \tag{4.33}$$

uniformly over z, where $j = 1$ or 2 depending on the symmetry of T_n. Usually $j = 1$ if T_n is a statistic for a one-tailed test and $j = 2$ if T_n is a statistic for a symmetrical, two-tailed test. $G_n(z, \psi^*)$ has an expansion identical to Equation (4.33) through $O(n^{-j/2})$. Therefore, through $O_p(n^{-j/2})$, bootstrap sampling when H_0 is false is equivalent to generating data from a model that satisfies H_0 with pseudo-true values of the parameters not specified by H_0. It follows that when H_0 is false, bootstrap-based critical values are equivalent through $O_p(n^{-j/2})$ to the critical values that would be obtained if the model satisfying H_0 with pseudo-true parameter values were correct. Moreover, the power of a test of H_0 using a bootstrap-based critical value is equal through $O(n^{-j/2})$ to the power against the true data-generation process that would be obtained by using the exact finite-sample critical value for testing H_0 with pseudo-true parameter values.

5. Monte Carlo experiments

This section presents the results of some Monte Carlo experiments that illustrate the numerical performance of the bootstrap as a means of reducing differences between the true and nominal rejection probabilities of tests of statistical hypotheses.

5.1. The information-matrix test

White's (1982) information-matrix (IM) test is a specification test for parametric models estimated by maximum likelihood. It tests the hypothesis that the Hessian and outer-product forms of the information matrix are equal. Rejection implies that the model is misspecified. The test statistic is asymptotically chi-square distributed, but Monte Carlo experiments carried out by many investigators have shown that the asymptotic distribution is a very poor approximation to the true, finite-sample distribution. With sample sizes in the range found in applications, the true and nominal probabilities that the IM test with asymptotic critical values rejects a correct model can differ by a factor of 10 or more [Horowitz (1994), Kennan and Neumann (1988), Orme (1990), Taylor (1987)].

Horowitz (1994) reports the results of Monte Carlo experiments that investigate the ability of the bootstrap to provide improved finite-sample critical values for the IM test, thereby reducing the distortions of RP's that occur with asymptotic critical values. Three forms of the test were used: the Chesher (1983) and Lancaster (1984) form, White's (1982) original form, and Orme's (1990) ω_3. The Chesher–Lancaster form is relatively easy to compute because, in contrast to the other forms, it does not require third derivatives of the log-density function or analytic expected values of derivatives of the log-density. However, first-order asymptotic theory gives an especially poor approximation to its finite-sample distribution. Orme (1990) found through Monte Carlo experimentation that the distortions of RP's are smaller with ω_3 than with many other forms of the IM test statistic. Orme's ω_3 uses expected values of third derivatives of the log-density, however, so it is relatively difficult to compute.

Table 1
Empirical rejection probabilities of nominal 0.05-level information-matrix tests of probit and tobit models[1]

n	Distribution of X	RP using asymptotic critical values			RP using bootstrap-based critical values		
		White	Chesh.-Lan.	Orme	White	Chesh.-Lan.	Orme
Binary probit models							
50	$N(0,1)$	0.385	0.904	0.006	0.064	0.056	0.033
	$U(-2,2)$	0.498	0.920	0.017	0.066	0.036	0.031
100	$N(0,1)$	0.589	0.848	0.007	0.053	0.059	0.054
	$U(-2,2)$	0.632	0.875	0.027	0.058	0.056	0.049
Tobit models							
50	$N(0,1)$	0.112	0.575	0.038	0.083	0.047	0.045
	$U(-2,2)$	0.128	0.737	0.174	0.051	0.059	0.054
100	$N(0,1)$	0.065	0.470	0.167	0.038	0.039	0.047
	$U(-2,2)$	0.090	0.501	0.163	0.046	0.052	0.039

[1] Source: Horowitz (1994).

Horowitz's (1994) experiments consisted of applying the three forms of the IM test to Tobit and binary probit models. Each model had either one or two explanatory variables X that were obtained by sampling either the $N(0,1)$ or the $U[0,1]$ distribution. There were 1000 replications in each experiment. Other details of the Monte Carlo procedure are described in Horowitz (1994). Table 1 summarizes the results of the experiments. As expected, the differences between empirical and nominal RP's are very large when asymptotic critical values are used. This is especially true for the Chesher–Lancaster form of the test. When bootstrap critical values are used, however, the differences between empirical and nominal RP's are very small. The bootstrap essentially eliminates the distortions of the RP's of the three forms of the IM test.

5.2. The t test in a heteroskedastic regression model

In this section, the heteroskedasticity-consistent covariance matrix estimator (HCCME) of Eicker (1963, 1967) and White (1980) is used to carry out a t test of a hypothesis about β in the model

$$Y = X\beta + U. \tag{5.1}$$

In this model, U is an unobserved random variable whose probability distribution is unknown and that may have heteroskedasticity of unknown form. It is assumed that $E(U|X=x)=0$ and $\mathrm{Var}(U|X=x)<\infty$ for all x in the support of X.

Let b_n be the ordinary least-squares (OLS) estimator of β in Equation (5.1), b_{ni} and β_i be the ith components of b_n and β, and s_{ni} be the square root of the (i, i) element of the HCCME. The t statistic for testing H_0: $\beta_i=\beta_{i0}$ is $T_n=(b_{ni}-\beta_{i0})/s_{ni}$. Under regularity conditions, $T_n \xrightarrow{d} N(0,1)$ as $n\rightarrow\infty$. However, Chesher and Jewitt (1987) have shown that s_{ni}^2 can be seriously biased downward. Therefore, the true RP of a test based on T_n is likely to exceed the nominal RP. As is shown later in this section, the differences between the true and nominal RP's can be very large when n is small.

The bootstrap can be implemented for model (5.1) by sampling observations of (Y,X) randomly with replacement. The resulting bootstrap sample is used to estimate β by OLS and compute T_n^*, the t statistic for testing H_0^*: $\beta_i=b_{ni}$. The empirical distribution of T_n^* is obtained by repeating this process many times, and the α-level bootstrap critical value for T_n^* is estimated from this distribution. Since U may be heteroskedastic, the bootstrap cannot be implemented by resampling OLS residuals independently of X. Similarly, one cannot implement the bootstrap by sampling U from a parametric model because Equation (5.1) does not specify the distribution of U or the form of any heteroskedasticity.

Randomly resampling (Y,X) pairs does not impose the restriction $E(U|X=x)=0$ on the bootstrap sample. As will be seen later in this section, the numerical performance of the bootstrap can be improved greatly through the use of an alternative resampling procedure, called the *wild bootstrap,* that imposes this restriction. The wild bootstrap was introduced by Liu (1988) following a suggestion of Wu (1986). Mammen (1993) establishes the ability of the wild bootstrap to provide asymptotic refinements for the model (5.1). Cao-Abad (1991), Härdle and Mammen (1993), and Härdle and Marron (1991) use the wild bootstrap in nonparametric regression.

To describe the wild bootstrap, write the estimated form of Equation (5.1) as

$$Y_i = X_i b_n + U_{ni}; \qquad i = 1,\ 2,\ \ldots,\ n,$$

where Y_i and X_i are the ith observed values of Y and X, and U_{ni} is the ith OLS residual. For each $i = 1, \ldots, n$, let F_i be the unique 2-point distribution that satisfies $E(Z|F_i) = 0$, $E(Z^2|F_i) = U_{ni}^2$, and $E(Z^3|F_i) = U_{ni}^3$, where Z is a random variable with the CDF F_i. Then, $Z = (1 - \sqrt{5})U_{ni}/2$ with probability $(1 + \sqrt{5})/(2\sqrt{5})$, and $Z = (1 + \sqrt{5})U_{ni}/2$ with probability $1 - (1 + \sqrt{5})/(2\sqrt{5})$. The wild bootstrap is implemented as follows:

(1) For each $i = 1, \ldots, n$, sample U_i^* randomly from F_i. Set $Y_i^* = X_i b_n + U_i^*$.

(2) Estimate Equation (5.1) by OLS using the bootstrap sample $\{Y_i^*, X_i{:}\ i = 1, \ldots, n\}$. Compute the resulting t statistic, T_n^*.

(3) Obtain the empirical distribution of the wild-bootstrap version of T_n^* by repeating steps 1 and 2 many times. Obtain the wild-bootstrap critical value of T_n^* from the empirical distribution.

Horowitz (1997) reports the results of a Monte Carlo investigation of the ability of the bootstrap and wild bootstrap to reduce the distortions in the RP of a symmetrical, two-tailed t test that occur when asymptotic critical values are used. The bootstrap was implemented by resampling (Y, X) pairs, and the wild bootstrap was implemented as described above. The experiments also investigate the RP of the t test when the HCCME is used with asymptotic critical values and when a jackknife version of the HCCME is used with asymptotic critical values [MacKinnon and White (1985)]. MacKinnon and White (1985) found through Monte Carlo experimentation that with the jackknife HCCME and asymptotic critical values, the t test had smaller distortions of RP than it did with several other versions of the HCCME.

The experiments use $n = 25$. X consists of an intercept and either 1 or 2 explanatory variables. In experiments in which X has an intercept and one explanatory variable, $\beta = (1, 0)'$. In experiments in which X has an intercept and two explanatory variables, $\beta = (1, 0, 1)'$. The hypothesis tested in all experiments is H_0: $\beta_2 = 0$. The components of X were obtained by independent sampling from a mixture of normal distributions in which $N(0, 1)$ was sampled with probability 0.9 and $N(2, 9)$ was sampled with probability 0.1. The resulting distribution of X is skewed and leptokurtotic. Experiments were carried out using homoskedastic and heteroskedastic U's. When U was homoskedastic, it was sampled randomly from $N(0, 1)$. When U was heteroskedastic, the U value corresponding to $X = x$ was sampled from $N(0, \Omega_x)$, where $\Omega_x = 1 + x^2$ or $\Omega_x = 1 + x_1^2 + x_2^2$, depending on whether X consists of 1 or 2 components in addition to an intercept. Ω_x is the covariance matrix of U corresponding to the

Table 2
Empirical rejection probabilities of t tests using heteroskedasticity-consistent covariance matrix estimators[1,2] ($n = 25$)

Form of test	1-Variable homoskedastic model	1-Variable random coeff. model	2-Variable homoskedastic model	2-Variable random coeff. model
Asymptotic	0.156	0.306	0.192	0.441
Jackknife	0.096	0.140	0.081	0.186
Bootstrap (Y, X) pairs	0.100	0.103	0.114	0.124
Wild bootstrap	0.050	0.034	0.062	0.057

[1] Source: Horowitz (1997).
[2] Empirical RP at nominal 0.05 level.

random-coefficients model $Y = X\beta + X\delta + V$, where V and the components of δ are independently distributed as $N(0, 1)$. There were 1000 Monte Carlo replications in each experiment.

Table 2 shows the empirical RP's of nominal 0.05-level t tests of H_0. The differences between the empirical and nominal RP's using the HCCME and asymptotic critical values are very large. Using the jackknife version of the HCCME or critical values obtained from the bootstrap greatly reduces the differences between the empirical and nominal RP's, but the empirical RP's are still 2–3 times the nominal ones. With critical values obtained from the wild bootstrap, the differences between the empirical and nominal RP's are very small. In these experiments, the wild bootstrap essentially removes the distortions of RP that occur with asymptotic critical values.

5.3. The t test in a Box–Cox regression model

The t statistic for testing a hypothesis about a slope coefficient in a linear regression model with a Box–Cox (1964) transformed dependent variable is not invariant to changes in the measurement units, or scale, of the dependent variable [Spitzer (1984)]. The numerical value of the t statistic and the finite-sample RP's of the t test with asymptotic critical values vary according to the measurement units or scale that is used. As a result, the finite-sample RP's of the t test with asymptotic critical values can be far from the nominal RP's. The bootstrap provides a better approximation to the finite-sample distribution and, therefore, better finite-sample critical values.

Horowitz (1997) reports the results of a Monte Carlo investigation of the finite-sample RP of a symmetrical t test of a hypothesis about a slope coefficient in a linear regression model with a Box–Cox transformed dependent variable. The model generating the data is

$$Y^{(\lambda)} = \beta_0 + \beta_1 X + U,$$

Table 3
Empirical rejection probabilities of t tests for Box–Cox regression model[1] (nominal RP = 0.05)

n	λ	Scale factor	RP using critical values from		Empirical critical values	Bootstrap critical values
			Asymptotic	Bootstrap		
50	0.01	0.2	0.048	0.066	1.930	1.860
		1.0	0.000	0.044	0.911	0.909
		5.0	0.000	0.055	0.587	0.571
100	0.01	0.2	0.047	0.053	1.913	1.894
		1.0	0.000	0.070	1.201	1.165
		5.0	0.000	0.056	0.767	0.759
50	1.0	0.2	0.000	0.057	1.132	1.103
		1.0	0.000	0.037	0.625	0.633
		5.0	0.000	0.036	0.289	0.287
100	1.0	0.2	0.000	0.051	1.364	1.357
		1.0	0.000	0.044	0.836	0.835
		5.0	0.000	0.039	0.401	0.391

[1] Source: Horowitz (1997).

where $Y^{(\lambda)}$ is the Box–Cox transformed value of the dependent variable Y, $U \sim N(0, \sigma^2)$, $\beta_0 = 2$, $\beta_1 = 0$ and $\sigma^2 = 0.0625$. X was sampled from $N(4, 4)$ and was fixed in repeated samples. The hypothesis being tested is H_0: $\beta_1 = 0$. The value of λ is either 0.01 or 1, depending on the experiment, and the scale of Y was 0.2, 1, or 5. The sample sizes were $n = 50$ and 100. There were 1000 replications in each experiment.

The results of the experiments are summarized in Table 3. The empirical critical value of the t test tends to be much smaller than the asymptotic critical value of 1.96, especially in the experiments with a scale factor of 5. As a result, the empirical RP of the t test is usually much smaller than its nominal RP. The mean bootstrap critical values, however, are very close to the empirical critical values, and the RP's based on bootstrap critical values are very close to the nominal ones.

5.4. *Estimation of covariance structures*

In estimation of covariance structures, the objective is to estimate the covariance matrix of a $k \times 1$ vector X subject to restrictions that reduce the number of unique, unknown elements to $r < k(k+1)/2$. Estimates of the r unknown elements can be obtained by minimizing the weighted distance between sample moments and the estimated population moments. Weighting all sample moments equally produces the equally-weighted minimum distance (EWMD) estimator, whereas choosing the weights to maximize asymptotic estimation efficiency produces the optimal minimum distance (OMD) estimator.

The OMD estimator dominates the EWMD estimator in terms of asymptotic efficiency, but it has been found to have poor finite-sample properties in applications [Abowd and Card (1989)]. Altonji and Segal (1994, 1996) carried out an extensive Monte Carlo investigation of the finite-sample performance of the OMD estimator. They found that the estimator is badly biased with samples of the sizes often found in applications and that its finite-sample root-mean-square estimation error (RMSE) often greatly exceeds the RMSE of the asymptotically inefficient EWMD estimator. Altonji and Segal also found that the true coverage probabilities of asymptotic confidence intervals based on the OMD estimator tend to be much lower than the nominal coverage probabilities. Thus, estimation and inference based on the OMD estimator can be highly misleading with finite samples.

Horowitz (1998a) reports the results of a Monte Carlo investigation the ability of the bootstrap to reduce the bias and RMSE of the OMD estimator and reduce the differences between true and nominal coverage probabilities of nominal 95% confidence intervals based on this estimator. The data-generation processes used in the Monte Carlo experiments were taken from Altonji and Segal (1994). In each experiment, X has 10 components, and the sample size is $n = 500$. The jth component of X, X_j $(j = 1, \dots, 10)$ is generated by $X_j = (Z_j + \rho Z_{j+1})/(1 + \rho^2)^{1/2}$, where $Z_1, \dots, Z_{11}$ are i.i.d. random variables with means of 0 and variances of 1, and $\rho = 0.5$. The Z's are sampled from five different distributions depending on the experiment. These are $U[0, 1]$, $N(0, 1)$, Student t with 10 degrees of freedom, exponential, and lognormal. It is assumed that ρ is known and that the components of X are known to be identically distributed and to follow MA(1) processes. The estimation problem is to infer the scalar parameter θ that is identified by the moment conditions $\text{Var}(X_j) = \theta$ $(j = 1, \dots, 10)$ and $\text{Cov}(X_j, X_{j-1}) = \rho\theta/(1 + \rho^2)$ $(j = 2, \dots, 10)$. Experiments were carried out with the EWMD and OMD estimators as well as a version of the OMD estimator that uses a trimmed estimator of the asymptotically optimal weight matrix. See Horowitz (1998a) for an explanation of the trimming procedure.

The results of the experiments are summarized in Table 4. The OMD estimator, $\theta_{n,\text{OMD}}$ is biased and its RMSE exceeds that of the EWMD estimator, $\theta_{n,\text{EWMD}}$ for all distributions of Z except the uniform. Moreover, the coverage probabilities of confidence intervals based on $\theta_{n,\text{OMD}}$ with asymptotic critical values are far below the nominal value of 0.95 except in the experiment with uniform Z's. Bootstrap bias reduction greatly reduces both the bias and RMSE of $\theta_{n,\text{OMD}}$. In addition, the use of bootstrap critical values greatly reduces the errors in the coverage probabilities of confidence intervals based on $\theta_{n,\text{OMD}}$. In the experiments with normal, Student t, or uniform Z's, the bootstrap essentially eliminates the bias of $\theta_{n,\text{OMD}}$ and the errors in the coverage probabilities of the confidence intervals. Moreover, the RMSE of the bias-corrected $\theta_{n,\text{OMD}}$ in these experiments is 12–50% less than that of $\theta_{n,\text{EWMD}}$.

When Z is exponential or lognormal, the bootstrap reduces but does not eliminate the bias of $\theta_{n,\text{OMD}}$ and the errors in the coverage probabilities of confidence intervals. Horowitz (1998a) shows that the poor performance of the bootstrap in these cases is caused by imprecise estimation of the OMD weight and covariance matrices. This

Table 4
Results of Monte Carlo experiments with estimators of covariance structures[1,2]

Dist.	EWMD	OMD without bootstrap			OMD with bootstrap			Trimmed OMD[5]		
	RMSE	Bias	RMSE	Cov.[3]	Bias	RMSE	Cov.[4]	Bias	RMSE	Cov.[4]
Uniform	0.019	0.005	0.015	0.93	0.002	0.014	0.96			
Normal	0.024	0.016	0.025	0.85	0.0	0.021	0.95			
Student *t*	0.029	0.024	0.034	0.79	0.002	0.026	0.95			
Exponential	0.042	0.061	0.073	0.54	0.014	0.048	0.91	0.004	0.042	0.96
Lognormal	0.138	0.274	0.285	0.03	0.136	0.173	0.76	0.046	0.126	0.91

[1] Source: Horowitz (1998a); nominal coverage probability is 0.95; based on 1000 replications.
[2] Abbreviations: Dist., distribution; EWMD, equally-weighted minimum distance; OMD, optimal minimum distance; RMSE, root-mean-square estimation error.
[3] Coverage probability with asymptotic critical value.
[4] Coverage probability with bootstrap critical value.
[5] Trimmed OMD with bootstrap.

problem is largely eliminated through the use of the trimmed estimator of these matrices. With trimming, $\theta_{n,\mathrm{OMD}}$ with exponential or lognormal Z's has a RMSE that is the same as or less than that of the EWMD estimator, and the empirical coverage probabilities of confidence intervals are close to the nominal values.

6. Conclusions

The bootstrap consistently estimates the asymptotic distributions of econometric estimators and test statistics under conditions that are sufficiently general to accommodate most applications. Subsampling methods usually can be used in place of the standard bootstrap when the latter is not consistent. Together, the bootstrap and subsampling methods provide ways to substitute computation for mathematical analysis if analytical calculation of the asymptotic distribution of an estimator or test statistic is difficult or impossible.

Under conditions that are stronger than those required for consistency but still general enough to accommodate a wide variety of econometric applications, the bootstrap reduces the finite-sample biases of estimators and provides a better approximation to the finite-sample distribution of an estimator or test statistic than does first-order asymptotic theory. The approximations of first-order asymptotic theory are often quite inaccurate with samples of the sizes encountered in applications. As a result, the true and nominal probabilities that a test rejects a correct hypothesis can be very different when critical values based on first-order approximations are used. Similarly, the true and nominal coverage probabilities of confidence intervals based on asymptotic critical values can be very different. The bootstrap can provide dramatic reductions in the differences between true and nominal rejection and coverage probabilities of tests and confidence intervals. In many cases of practical importance, the bootstrap essentially eliminates finite-sample errors in rejection and coverage probabilities.

This chapter has also emphasized the need for care in applying the bootstrap. The importance of asymptotically pivotal statistics for obtaining asymptotic refinements has been stressed. Proper attention also must be given to matters such as recentering, correction of test statistics in the block bootstrap for dependent data, smoothing, and choosing the distribution from which bootstrap samples are drawn. These qualifications do not, however, detract from the importance of the bootstrap as a practical tool for improving inference in applied econometrics.

Acknowledgements

I thank Donald Andrews, James Heckman, Hidehiko Ichimura, John Kennan, Edward Leamer, Catherine Loader, George Neumann, Harry Paarsch, Bernard Salanié, and Gene Savin for comments on earlier drafts of this chapter. Parts of the chapter are taken from Horowitz (1997) and are used here with permission from the Cambridge University Press. The preparation of this chapter was supported in part by NSF Grants SBR-9617925 and SES 9910925.

Appendix A. Informal derivation of Equation (3.27)

To derive Equation (3.27), write $P(|T_n| \geqslant z^*_{n,\alpha/2})$ in the form

$$\begin{aligned} P(|T_n| > z^*_{n,\alpha/2}) &= 1 - [P(T_n \leqslant z^*_{n,\alpha/2}) - P(T_n \leqslant -z^*_{n,\alpha/2})] \\ &= 1 - \{P[T_n - (z^*_{n,\alpha/2} - z_{\infty,\alpha/2}) \leqslant z_{\infty,\alpha/2}] \\ &\qquad - P[T_n + (z^*_{n,\alpha/2} - z_{\infty,\alpha/2}) \leqslant -z_{\infty,\alpha/2}]\}. \end{aligned} \tag{A.1}$$

With an error whose size is almost surely $O(n^{-2})$, $(z^*_{n,\alpha/2} - z_{\infty,\alpha/2})$ on the right-hand side of (A.1) can be replaced with a Cornish–Fisher expansion that retains terms through $O(n^{-3/2})$. This expansion can be obtained by applying the delta method to the difference between Equations (3.23) and (3.24). The result is

$$z^*_{n,\alpha/2} - z_{\infty,\alpha/2} = -\frac{1}{n}\frac{g_2(z_{\infty,\alpha/2}, F_0)}{\phi(z_{\infty,\alpha/2})} + \frac{1}{n^{3/2}} n^{1/2} r_3(\bar{Z}) + O(n^{-2}), \tag{A.2}$$

where r_3 is a smooth function, $r_3(\mu_Z) = 0$, and $n^{1/2} r_3(\bar{Z}) = O_p(1)$ as $n \to \infty$. Substituting Equation (A.2) into Equation (A.1) yields

$$\begin{aligned} P(|T_n| > z^*_{n,\alpha/2}) = 1 - \{&P[T_n - n^{-3/2} n^{1/2} r_3(\bar{Z}) \leqslant z_{\infty,\alpha/2} + n^{-1} r_2(z_{\infty,\alpha/2})] \\ &- P[T_n + n^{-3/2} n^{1/2} r_3(\bar{Z}) \leqslant -z_{\infty,\alpha/2} - n^{-1} r_2(z_{\infty,\alpha/2})]\} + O(n^{-2}), \end{aligned} \tag{A.3}$$

where

$$r_2(z) = -\frac{g_2(z, F_0)}{\phi(z)}. \tag{A.4}$$

The next step is to replace the right-hand side of Equation (A.3) with an Edgeworth approximation. To do this, it is necessary to provide a detailed specification of the function g_2 in Equations (3.9) and (3.13). Let $\kappa_{j,n}$ denote the jth cumulant of T_n[23]. Under assumption SFM, $\kappa_{j,n}$ can be expanded in a power series. For a statistic such as T_n whose asymptotic distribution has a variance of 1,

$$\kappa_{1,n} = \frac{k_{12}}{n^{1/2}} + \frac{k_{13}}{n^{3/2}} + \mathrm{O}(n^{-5/2}),$$

$$\kappa_{2,n} = 1 + \frac{k_{22}}{n} + \mathrm{O}(n^{-2}),$$

$$\kappa_{3,n} = \frac{k_{31}}{n^{1/2}} + \frac{k_{32}}{n^{3/2}} + \mathrm{O}(n^{-5/2}),$$

and

$$\kappa_{4,n} = \frac{k_{41}}{n} + \mathrm{O}(n^{-2}),$$

where the coefficients k_{jk} are functions of moments of products of components of Z. The function g_2 is then

$$g_2(\tau, F_0) = -\tau\left[\tfrac{1}{2}(k_{22} + k_{12}^2) + \tfrac{1}{24}(k_{41} + 4k_{12}k_{31})(\tau^2 - 3) + \tfrac{1}{72}k_{31}^2(\tau^4 - 10\tau^2 + 15)\right]\phi(\tau). \tag{A.5}$$

See Hall (1992a, pp. 46–56) for details. Denote the quantity on the right-hand side of Equation (A.5) by $\tilde{g}_2(\tau, \kappa_0)$, where κ_0 denotes the k_{jk} coefficients that are associated with cumulants of the distribution of T_n. Let $\hat{\kappa}_n$ denote the k_{jk} coefficients that are associated with cumulants of $T_n \pm n^{-3/2}n^{1/2}r_3(\bar{Z})$, and let $\tilde{g}_2(\tau, \hat{\kappa}_n)$ denote the version of $\tilde{g}_2$ that is obtained by replacing κ_0 with $\hat{\kappa}_n$. The difference between the + and – coefficients is asymptotically negligible. Now replace $g_2(\tau, F_0)$ in Equation (3.13) with $\tilde{g}_2(\tau, \hat{\kappa}_n)$. Also, replace τ with $z_{\infty,\alpha/2} + n^{-1}r_2(z_{\infty,\alpha/2})$ in Equation (3.13). Substituting the result into the right-hand side of Equation (A.3) gives the following Edgeworth approximation to $P(|T_n| > z^*_{n,\alpha/2})$:

$$\begin{aligned} P(|T_n| > z^*_{n,\alpha/2}) = 2\{1 - \Phi[z_{\infty,\alpha/2} + n^{-1}r_2(z_{\infty,\alpha/2})]\} \\ - 2n^{-1}\tilde{g}_2[z_{\infty,\alpha/2} + n^{-1}r_2(z_{\infty,\alpha/2}), \hat{\kappa}_n] + \mathrm{O}(n^{-2}). \end{aligned} \tag{A.6}$$

A Taylor-series expansion of the right-hand side of Equation (A.6) combined with Equation (A.4) and the fact that $2[1 - \Phi(z_{\infty,\alpha/2})] = \alpha$ gives

$$P(|T_n| > z^*_{n,\alpha/2}) = \alpha + \frac{2}{n}[\tilde{g}_2(z_{\infty,\alpha/2}, \kappa_0) - \tilde{g}_2(z_{\infty.\alpha/2}, \hat{\kappa}_n)] + \mathrm{O}(n^{-2}). \tag{A.7}$$

It is not difficult to show that $\tilde{g}_2(z_{\infty,\alpha/2}, \kappa_0) - \tilde{g}_2(z_{\infty,\alpha/2}, \hat{\kappa}_n) = \mathrm{o}(n^{-1})$. (Roughly speaking, this is because $n^{-1}r_3(\bar{Z}) = \mathrm{o}(n^{-1})$ almost surely.) Therefore, the second term on the right-hand side of Equation (A.7) is $\mathrm{o}(n^{-2})$, which yields Equation (3.27).

[23] The cumulants of a distribution are coefficients in a power-series expansion of the logarithm of its characteristic function. The first three cumulants are the mean, variance, and third moment about the mean. The fourth cumulant is the fourth moment about the mean minus three times the square of the variance.

References

Abowd, J.M., and D. Card (1987), "Intertemporal labor supply and long-term employment contracts", American Economic Review 77:50–68.

Abowd, J.M., and D. Card (1989), "On the covariance of earnings and hours changes", Econometrica 57:411–445.

Altonji, J.G., and L.M. Segal (1994), "Small sample bias in GMM estimation of covariance structures", NBER Technical Working Paper no. 156 (National Bureau of Economic Research, Cambridge, MA).

Altonji, J.G., and L.M. Segal (1996), "Small sample bias in GMM estimation of covariance structures", Journal of Business and Economic Statistics 14:353–366.

Amemiya, T. (1985), Advanced Econometrics (Harvard University Press, Cambridge, MA).

Amemiya, T., and J.L. Powell (1981), "A comparison of the Box–Cox maximum likelihood estimator and the non-linear two-stage least squares estimator", Journal of Econometrics 17:351–381.

Andrews, D.W.K. (1991), "Heteroskedasticity and autocorrelation consistent covariance matrix estimation", Econometrica 59:817–858.

Andrews, D.W.K. (1997), "A conditional Kolmogorov test", Econometrica 65:1097–1128.

Andrews, D.W.K. (1999), "Higher-order improvements of a computationally attractive k-step bootstrap for extremum estimators", Cowles Foundation discussion paper No. 1230 (Cowles Foundation for Research in Economics, Yale University).

Andrews, D.W.K. (2000), "Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space", Econometrica 68:399–405.

Andrews, D.W.K., and M. Buchinsky (2000), "A three-step method for choosing the number of bootstrap repetitions", Econometrica 68:23–51.

Andrews, D.W.K., and J.C. Monahan (1992), "An improved heteroskedasticity and autocorrelation consistent covariance matrix", Econometrica 59:817–858.

Athreya, K. (1987), "Bootstrap of the mean in the infinite variance case", Annals of Statistics 15:724–731.

Babu, G.J., and K. Singh (1983), "Inference on means using the bootstrap", Annals of Statistics 11:999–1003.

Babu, G.J., and K. Singh (1984), "On one term correction by Efron's bootstrap", Sankhya Series A 46:219–232.

Basawa, I.V., A.K. Mallik, W.P. McCormick, J.H. Reeves and R.L. Taylor (1991a), "Bootstrapping unstable first-order autoregressive processes", Annals of Statistics 19:1098–1101.

Basawa, I.V., A.K. Mallik, W.P. McCormick, J.H. Reeves and R.L. Taylor (1991b), "Bootstrap test of significance and sequential bootstrap estimation for unstable first order autoregressive processes", Communications in Statistics – Theory and Methods 20:1015–1026.

Bassett, G., and R. Koenker (1978), "Asymptotic theory of least absolute error regression", Journal of the American Statistical Association 73:618–621.

Behrman, J.R., M.R. Rosenzweig and P. Taubman (1994), "Endowments and the allocation of schooling in the family and in the marriage market: the twins experiment", Journal of Political Economy 102:1131–1174.

Beran, R. (1982), "Estimated sampling distributions: The bootstrap and competitors", Annals of Statistics 10:212–225.

Beran, R. (1987), "Prepivoting to reduce level error of confidence sets", Biometrika 74:457–468.

Beran, R. (1988), "Prepivoting test statistics: a bootstrap view of asymptotic refinements", Journal of the American Statistical Association 83:687–697.

Beran, R., and G.R. Ducharme (1991), Asymptotic Theory for Bootstrap Methods in Statistics (Les Publications CRM, Centre de recherches mathématiques, Université de Montréal, Montréal, Canada).

Bertail, P., D.N. Politis and J.P. Romano (1999), "On subsampling estimators with unknown rate of convergence", Journal of the American Statistical Association 94:569–579.

Bickel, P.J., and D.A. Freedman (1981), "Some asymptotic theory for the bootstrap", Annals of Statistics 9:1196–1217.

Bickel, P.J., F. Götze and W.R. van Zwet (1997), "Resampling fewer than *n* observations: gains, losses, and remedies for losses", Statistica Sinica 7:1–32.

Blanchard, O.J., and D. Quah (1989), "The dynamic effects of aggregate demand and supply disturbances", American Economic Review 79:655–673.

Bose, A. (1988), "Edgeworth correction by bootstrap in autoregressions", Annals of Statistics 16: 1709–1722.

Bose, A. (1990), "Bootstrap in moving average models", Annals of the Institute of Statistical Mathematics 42:753–768.

Box, G.E.P., and D.R. Cox (1964), "An analysis of transformations", Journal of the Royal Statistical Society, Series B 26:211–243.

Brown, B., W.K. Newey and S. May (1997), "Efficient bootstrapping for GMM", Unpublished manuscript (Department of Economics, Massachusetts Institute of Technology).

Brown, B.W. (1999), "Simulation variance reduction for bootstrapping", in: R. Mariano, T. Schuermann, and M. Weeks, eds., Simulation-Based Econometrics: Methods and Applications (Cambridge University Press, New York).

Bühlmann, P. (1997), "Sieve bootstrap for time series", Bernoulli 3:123–148.

Bühlmann, P. (1998), "Sieve bootstrap for smoothing in nonstationary time series", Annals of Statistics 26:48–83.

Cao-Abad, R. (1991), "Rate of convergence for the wild bootstrap in nonparametric regression", Annals of Statistics 19:2226–2231.

Carlstein, E. (1986), "The use of subseries methods for estimating the variance of a general statistic from a stationary time series", Annals of Statistics 14:1171–1179.

Cavanagh, C.L. (1987), "Limiting behavior of estimators defined by optimization", Unpublished manuscript (Department of Economics, Harvard University).

Chandra, T.K., and J.K. Ghosh (1979), "Valid asymptotic expansions for the likelihood ratio statistic and other perturbed chi-square variables", Sankhya Series A 41:22–47.

Chesher, A. (1983), "The information matrix test", Economics Letters 13:45–48.

Chesher, A., and I. Jewitt (1987), "The bias of a heteroskedasticity consistent covariance matrix estimator", Econometrica 55:1217–1222.

Choi, E., and P. Hall (2000), "Bootstrap confidence regions computed from autoregressions of arbitrary order", Journal of the Royal Statistical Society, Series B 62:461–477.

Datta, S. (1995), "On a modified bootstrap for certain asymptotically non-normal statistics", Statistics and Probability Letters 24:91–98.

Datta, S. (1996), "On asymptotic properties of bootstrap for AR(1) processes", Journal of Statistical Planning and Inference 53:361–374.

Datta, S., and W.P. McCormick (1995), "Some continuous Edgeworth expansions for Markov chains with applications to bootstrap", Journal of Multivariate Analysis 52:83–106.

Davidson, R., and J.G. MacKinnon (1999a), "Bootstrap testing in nonlinear models", International Economic Review 40:487–508.

Davidson, R., and J.G. MacKinnon (1999b), "The size distortion of bootstrap tests", Econometric Theory 15:361–376.

Davison, A.C., and D.V. Hinkley (1997), Bootstrap Methods and Their Application (Cambridge University Press, Cambridge, U.K).

De Angelis, D., P. Hall and G.A. Young (1993), "Analytical and bootstrap approximations to estimator distributions in L^1 regression", Journal of the American Statistical Association 88:1310–1316.

Donald, S.G., and H.J. Paarsch (1996), "Identification, estimation, and testing in empirical models of auctions within the independent private values paradigm", Econometric Theory 12:517–567.

Efron, B. (1979), "Bootstrap methods: another look at the jackknife", Annals of Statistics 7:1–26.

Efron, B. (1987), "Better bootstrap confidence intervals", Journal of the American Statistical Association 82:171–185.

Efron, B., and R.J. Tibshirani (1993), An Introduction to the Bootstrap (Chapman & Hall, New York).

Eicker, F. (1963), "Asymptotic normality and consistency of the least squares estimators for families of linear regressions", Annals of Mathematical Statistics 34:447–456.

Eicker, F. (1967), "Limit theorems for regression with unequal and dependent errors", in: L. LeCam and J. Neyman, eds., Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability (University of California Press, Berkeley, CA) 59-82.

Ferretti, N., and J. Romo (1996), "Unit root bootstrap tests for AR(1) models", Biometrika 83:849–860.

Flinn, C.J., and J.J. Heckman (1982), "New methods for analyzing structural models of labor force dynamics", Journal of Econometrics 18:115–168.

Freedman, D.A. (1981), "Bootstrapping regression models", Annals of Statistics 9:1218–1228.

Gill, R.D. (1989), "Non- and semi-parametric maximum likelihood estimators and the von Mises method (Part 1)", Scandinavian Journal of Statistics 16:97–128.

Götze, F., and C. Hipp (1983), "Asymptotic expansions for sums of weakly dependent random vectors", Zeitschrift für Warscheinlichkeitstheorie und verwandte Gebiete 64:211–239.

Götze, F., and C. Hipp (1994), "Asymptotic distribution of statistics in time series", Annals of Statistics 22:2062–2088.

Götze, F., and H.R. Künsch (1996), "Blockwise bootstrap for dependent observations: higher order approximations for studentized statistics", Annals of Statistics 24:1914–1933.

Griliches, Z. (1979), "Sibling models and data in economics: beginnings of a survey", Journal of Political Economy 87:S37–S64.

Hahn, J. (1995), "Bootstrapping the quantile regression estimators", Econometric Theory 11:105–121.

Hahn, J. (1996), "A note on bootstrapping generalized method of moments estimators", Econometric Theory 12:187–197.

Hall, P. (1985), "Resampling a coverage process", Stochastic Process Applications 19:259–269.

Hall, P. (1986a), "On the number of bootstrap simulations required to construct a confidence interval", Annals of Statistics 14:1453–1462.

Hall, P. (1986b), "On the bootstrap and confidence intervals", Annals of Statistics 14:1431–1452.

Hall, P. (1988), "Theoretical comparison of bootstrap confidence intervals", Annals of Statistics 16: 927–953.

Hall, P. (1990), "Asymptotic properties of the bootstrap for heavy-tailed distributions", Annals of Probability 18:1342–1360.

Hall, P. (1992a), The Bootstrap and Edgeworth Expansion (Springer, New York).

Hall, P. (1992b), "Effect of bias estimation on coverage accuracy of bootstrap confidence intervals for a probability density", Annals of Statistics 20:675–694.

Hall, P. (1994), "Methodology and theory for the bootstrap", in: R.F. Engle and D.F. McFadden, eds., Handbook of Econometrics, vol. 4 (Elsevier, Amsterdam).

Hall, P., and J.L. Horowitz (1996), "Bootstrap critical values for tests based on generalized-method-of-moments estimators", Econometrica 64:891–916.

Hall, P., and B.-Y. Jing (1996), "On sample reuse methods for dependent data", Journal of the Royal Statistical Society Series B 58:727–737.

Hall, P., J.L. Horowitz and B.-Y. Jing (1995), "On blocking rules for the bootstrap with dependent data", Biometrika 82:561–574.

Hall, R.E., and F.S. Mishkin (1982), "The sensitivity of consumption to transitive income: estimates from panel data on households", Econometrica 50:461–481.

Hansen, L.P. (1982), "Large sample properties of generalized method of moments estimators", Econometrica 50:1029–1054.

Hansen, L.P., and K. Singleton (1982), "Generalized instrumental variables estimation of nonlinear rational expectations models", Econometrica 50:1269–1286.

Härdle, W. (1990), Applied Nonparametric Regression (Cambridge University Press Cambridge, UK).

Härdle, W., and E. Mammen (1993), "Comparing nonparametric versus parametric regression fits", Annals of Statistics 21:1926–1947.

Härdle, W., and J.S. Marron (1991), "Bootstrap simultaneous error bars for nonparametric regression", Annals of Statistics 19:778–796.

Härdle, W., W. Hildenbrand and M. Jerison (1991), "Empirical evidence on the law of demand", Econometrica 59:1525–1550.

Härdle, W., S. Huet and E. Jolivet (1995), "Better bootstrap confidence intervals for regression curve estimation", Statistics 26:287–306.

Heckman, J.J., J. Smith and N. Clements (1997), "Making the most out of programme evaluations and social experiments: accounting for heterogeneity in programme impacts", Review of Economic Studies 64:487–535.

Hillier, G.H. (1985), "On the joint and marginal densities of instrumental variables estimators in a general structural equation", Econometric Theory 1:53–72.

Horowitz, J.L. (1992), "A smoothed maximum score estimator for the binary response model", Econometrica 60:505–531.

Horowitz, J.L. (1994), "Bootstrap-based critical values for the information-matrix test", Journal of Econometrics 61:395–411.

Horowitz, J.L. (1997), "Bootstrap methods in econometrics: theory and numerical performance", in: D.M. Kreps and K.F. Wallis, eds., Advances in Economics and Econometrics: Theory and Applications, Seventh World Congress, Vol. 3 (Cambridge University Press, Cambridge, UK).

Horowitz, J.L. (1998a), "Bootstrap methods for covariance structures", Journal of Human Resources 33:39–61.

Horowitz, J.L. (1998b), "Bootstrap methods for median regression models", Econometrica 66:1327–1351.

Horowitz, J.L. (1998c), "Bootstrap critical values for tests based on the smoothed maximum score estimator", Journal of Econometrics, forthcoming.

Janas, D. (1993), "A smoothed bootstrap estimator for a studentized sample quantile", Annals of the Institute of Statistical Mathematics 45:317–329.

Jeong, J., and G.S. Maddala (1993), "A perspective on application of bootstrap methods in econometrics", in: G.S. Maddala, C.R. Rao, and H.D. Vinod, eds., Handbook of Statistics, Vol. 11 (North-Holland, Amsterdam).

Kennan, J.F., and G.R. Neumann (1988), "Why does the information matrix test reject so often?" Working paper no. 88-4 (Department of Economics, University of Iowa).

Kim, J., and D. Pollard (1990), "Cube root asymptotics", Annals of Statistics 18:191–219.

Kitamura, Y. (1997), "Empirical likelihood methods with weakly dependent processes", Annals of Statistics 25:2084–2102.

Koenker, R., and G. Bassett (1978), "Regression quantiles", Econometrica 46:33–50.

Kreiss, J.-P. (1992), "Bootstrap procedures for AR(∞) processes", in: K.H. Jöckel, G. Rothe and W. Sender, eds., Bootstrapping and Related Techniques, Lecture Notes in Economics and Mathematical Systems 376 (Springer-Verlag, Heidelberg).

Künsch, H.R. (1989), "The jackknife and the bootstrap for general stationary observations", Annals of Statistics 17:1217–1241.

Lahiri, S. (1996), "On Edgeworth expansion and moving block bootstrap for Studentized M-estimators in multiple linear regression models", Journal of Multivariate Analysis 56:42–59.

Lahiri, S.N. (1992), "Edgeworth correction by 'moving block' bootstrap for stationary and nonstationary data", in: R. LePage and L. Billard, eds., Exploring the Limits of Bootstrap (Wiley, New York).

Lahiri, S.N. (1999), "Theoretical comparisons of block bootstrap methods", Annals of Statistics 27: 386–404.

Lancaster, T. (1984), "The covariance matrix of the information matrix test", Econometrica 52:1051–1053.

Lehmann, E.L. (1959), Testing Statistical Hypotheses (Wiley, New York).

Li, H., and G.S. Maddala (1996), "Bootstrapping time series models", Econometric Reviews 15:115–158.

Li, H., and G.S. Maddala (1997), "Bootstrapping cointegrating regressions", Journal of Econometrics 80:297–318.

Liu, R.Y. (1988), "Bootstrap procedures under some non-i.i.d. models", Annals of Statistics 16: 1696–1708.

Liu, R.Y., and K. Singh (1987), "On a partial correction by the bootstrap", Annals of Statistics 15:1713–1718.

MacKinnon, J.G., and H. White (1985), "Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties", Journal of Econometrics 29:305–325.

Mammen, E. (1992), When Does Bootstrap Work? Asymptotic Results and Simulations (Springer, New York).

Mammen, E. (1993), "Bootstrap and wild bootstrap for high dimensional linear models", Annals of Statistics 21:255–285.

Manski, C.F. (1975), "Maximum score estimation of the stochastic utility model of choice", Journal of Econometrics 3:205–228.

Manski, C.F. (1985), "Semiparametric analysis of discrete response: asymptotic properties of the maximum score estimator", Journal of Econometrics 27:313–334.

Manski, C.F., and T.S. Thompson (1986), "Operational characteristics of maximum score estimation", Journal of Econometrics 32:85–108.

Nelson, C.R., and R. Startz (1990a), "The distribution of the instrumental variable estimator and its *t* ratio when the instrument is a poor one", Journal of Business 63:S125–S140.

Nelson, C.R., and R. Startz (1990b), "Some further results on the exact small sample properties of the instrumental variable estimator", Econometrica 58:967–976.

Newey, W.K., and K.D. West (1987), "A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix", Econometrica 55:703–708.

Newey, W.K., and K.D. West (1994), "Automatic lag selection in covariance matrix estimation", Review of Economic Studies 61:631–653.

Orme, C. (1990), "The small-sample performance of the information-matrix test", Journal of Econometrics 46:309–331.

Paparoditis, E., and D.N. Politis (2000), "The local bootstrap for Markov processes", Journal of Statistical Planning and Inference, forthcoming.

Phillips, P.C.B. (1983), "Exact small sample theory in the simultaneous equations model", in: Z. Griliches and M.D. Intriligator, eds., Handbook of Econometrics, Vol. 1 (North-Holland, Amsterdam).

Politis, D.N., and J.P. Romano (1994), "Large sample confidence regions based on subsamples under minimal assumptions", Annals of Statistics 22:2031–2050.

Politis, D.N., J.P. Romano and M. Wolf (1997), "Subsampling for heteroskedastic time series", Journal of Econometrics 81:281–317.

Politis, D.N., J.P. Romano and M. Wolf (1999), Subsampling (Springer, New York).

Powell, J.L. (1984), "Least absolute deviations estimation for the censored regression model", Journal of Econometrics 25:303–325.

Powell, J.L. (1986), "Censored regression quantiles", Journal of Econometrics 32:143–155.

Rajarshi, M.B. (1990), "Bootstrap in Markov-sequences based on estimates of transition density", Annals of the Institute of Statistical Mathematics 42:253–268.

Rao, C.R. (1973), Linear Statistical Inference and its Applications, 2nd Edition. Wiley, New York.

Runkle, D.E. (1987), "Vector autoregressions and reality", Journal of Business and Economic Statistics 5:437–442.

Shao, U., and D. Tu (1995), The Jackknife and Bootstrap (Springer, New York).

Silverman, B.W. (1986), Density Estimation for Statistics and Data Analysis (Chapman and Hall, London).

Singh, K. (1981), "On the asymptotic accuracy of Efron's bootstrap", Annals of Statistics 9:1187–1195.

Spitzer, J.J. (1984), "Variance estimates in models with the Box–Cox transformation: implications for estimation and hypothesis testing", Review of Economics and Statistics 66:645–652.

Stute, W. (1997), “Nonparametric model checks for regression”, Annals of Statistics 25:613–641.

Swanepoel, J.W.H. (1986), “A note on proving that the (modified) bootstrap works”, Communications in Statistics Theory and Methods 15:3193–3203.

Taylor, L.W. (1987), “The size bias of White’s information matrix test”, Economics Letters 24:63–67.

Vinod, H.D. (1993), “Bootstrap methods: applications in econometrics”, in: G.S. Maddala, C.R. Rao and H.D. Vinod, eds., Handbook of Statistics, Vol. 11 (North-Holland, Amsterdam).

West, K.D. (1990), “The sources of fluctuations in aggregate inventories and GNP”, Quarterly Journal of Economics 105:939–971.

White, H. (1980), “A heteroscedasticity-consistent covariance matrix estimator and a direct test for heteroscedasticity”, Econometrica 48:817–838.

White, H. (1982), “Maximum likelihood estimation of misspecified models”, Econometrica 50:1–26.

Wu, C.F.J. (1986), “Jackknife, bootstrap and other resampling methods in regression analysis”, Annals of Statistics 14:1261–1295.

Chapter 53

PANEL DATA MODELS: SOME RECENT DEVELOPMENTS *

MANUEL ARELLANO

CEMFI, Casado del Alisal 5, 28014 Madrid, Spain

BO HONORÉ

Department of Economics, Princeton University, Princeton, New Jersey 08544

Contents

Abstract 3231
Keywords 3231
1. Introduction 3232
2. Linear models with predetermined variables: identification 3233
   2.1. Strict exogeneity, predeterminedness, and unobserved heterogeneity 3233
   2.2. Time series models with error components 3241
      2.2.1. The AR(1) process with fixed effects 3242
      2.2.2. Aggregate shocks 3245
      2.2.3. Identification and unit roots 3246
      2.2.4. The value of information with highly persistent data 3247
   2.3. Using stationarity restrictions 3248
   2.4. Models with multiplicative effects 3250
3. Linear models with predetermined variables: estimation 3255
   3.1. GMM estimation 3255
   3.2. Efficient estimation under conditional mean independence 3259
   3.3. Finite sample properties of GMM and alternative estimators 3262
   3.4. Approximating the distributions of GMM and LIML for AR(1) models when the number of moments is large 3264
4. Nonlinear panel data models 3265
5. Conditional maximum likelihood estimation 3267
   5.1. Conditional maximum likelihood estimation of logit models 3268
   5.2. Poisson regression models 3269
6. Discrete choice models with “fixed” effects 3270

* We thank Jason Abrevaya, Badi Baltagi, Olympia Bover, Martin Browning, Jim Heckman, Luojia Hu, Ekaterini Kyriazidou, Ed Leamer, Aprajit Mahajan, Enrique Sentana, Jeffrey Wooldridge and participants at the Chicago and London Handbook conferences for helpful comments. All errors are our responsibility.

Handbook of Econometrics, Volume 5, Edited by J.J. Heckman and E. Leamer

7. Tobit-type models with “fixed” effects 3272
7.1. Censored regression models 3272
7.2. Type 2 Tobit model (sample selection model) 3275
7.3. Other Tobit-type models 3276
7.4. Monotone transformation models 3278
7.5. Nonparametric regression and fixed effects 3280
7.6. Relationship with estimators for some cross sectional models 3280
8. Models with lagged dependent variables 3282
8.1. Discrete choice with state dependence 3283
8.2. Dynamic Tobit models 3286
8.3. Dynamic sample selection models 3287
9. “Random” effects models 3287
10. Concluding remarks 3290
References 3290

Abstract

This chapter focuses on two of the developments in panel data econometrics since the Handbook chapter by Chamberlain (1984).

The first objective of this chapter is to provide a review of linear panel data models with predetermined variables. We discuss the implications of assuming that explanatory variables are predetermined as opposed to strictly exogenous in dynamic structural equations with unobserved heterogeneity. We compare the identification from moment conditions in each case, and the implications of alternative feedback schemes for the time series properties of the errors. We next consider autoregressive error component models under various auxiliary assumptions. There is a trade-off between robustness and efficiency since assumptions of stationary initial conditions or time series homoskedasticity can be very informative, but estimators are not robust to their violation. We also discuss the identification problems that arise in models with predetermined variables and multiple effects. Concerning inference in linear models with predetermined variables, we discuss the form of optimal instruments, and the sampling properties of GMM and LIML-analogue estimators drawing on Monte Carlo results and asymptotic approximations.

A number of identification results for limited dependent variable models with fixed effects and strictly exogenous variables are available in the literature, as well as some results on consistent and asymptotically normal estimation of such models. There are also some results available for models of this type including lags of the dependent variable, although even less is known for nonlinear dynamic models. Reviewing the recent work on discrete choice and selectivity models with fixed effects is the second objective of this chapter. A feature of parametric limited dependent variable models is their fragility to auxiliary distributional assumptions. This situation prompted the development of a large literature dealing with semiparametric alternatives (reviewed in Powell, 1994's chapter). The work that we review in the second part of the chapter is thus at the intersection of the panel data literature and that on cross-sectional semiparametric limited dependent variable models.

Keywords

JEL classification: C33

1. Introduction

Panel data analysis is at the watershed of time series and cross-section econometrics. While the identification of time series parameters traditionally relied on notions of stationarity, predeterminedness and uncorrelated shocks, cross-sectional parameters appealed to exogenous instrumental variables and random sampling for identification. By combining the time series and cross-sectional dimensions, panel datasets have enriched the set of possible identification arrangements, and forced economists to think more carefully about the nature and sources of identification of parameters of potential interest.

One strand of the literature found its original motivation in the desire of exploiting panel data for controlling unobserved time-invariant heterogeneity in cross-sectional models. Another strand was interested in panel data as a way to disentangle components of variance and to estimate transition probabilities among states. Papers in these two veins can be loosely associated with the early work on fixed and random effects approaches, respectively. In the former, interest typically centers in measuring the effect of regressors holding unobserved heterogeneity constant. In the latter, the parameters of interest are those characterizing the distributions of the error components. A third strand of the literature studied autoregressive models with individual effects, and more generally models with lagged dependent variables.

A sizeable part of the work in the first two traditions concentrated on models with just strictly exogenous variables. This contrasts with the situation in time series econometrics where the distinction between predetermined and strictly exogenous variables has long been recognized as a fundamental one in the specification of empirical models.

The first objective of this chapter is to review recent work on linear panel data models with predetermined variables. Lack of control of individual heterogeneity could result in a *spurious* rejection of strict exogeneity, and so a definition of strict exogeneity conditional on unobserved individual effects is a useful extension of the standard concept to panel data (a major theme of Chamberlain, 1984's chapter). There are many instances, however, in which for theoretical or empirical reasons one is concerned with models exhibiting *genuine* lack of strict exogeneity after controlling for individual heterogeneity.

The interaction between unobserved heterogeneity and predetermined regressors in short panels – which are the typical ones in microeconometrics – poses identification problems that are absent from both time series models and panel data models with only strictly exogenous variables. In our review we shall see that for linear models it is possible to accommodate techniques developed from the various strands in a common framework within which their relative merits can be evaluated.

Much less is known for discrete choice, selectivity and other non-linear models of interest in microeconometrics. A number of identification results for limited dependent variable models with fixed effects and strictly exogenous variables are available in the literature, as well as some results on consistent and asymptotically normal estimation of such models. There are also some results available for models of this type including lags of the dependent variable, although even less is known for nonlinear dynamic models.

Reviewing the recent work on discrete choice and selectivity models with fixed effects is the second objective of this chapter. A feature of parametric limited dependent variable models is their fragility to auxiliary distributional assumptions. This situation prompted the development of a large literature dealing with semiparametric alternatives (reviewed in Powell, 1994's chapter). The work that we review in the second part of the chapter is thus at the intersection of the panel data literature and that on cross-sectional semiparametric limited dependent variable models.

Other interesting topics in panel data analysis which will not be covered in this chapter include work on long T panel data models with heterogeneous dynamics or unit roots [Pesaran and Smith (1995), Canova and Marcet (1995), Kao (1999), Phillips and Moon (1999)], simulation-based random effects approaches to the nonlinear models [Hajivassiliou and McFadden (1990), Keane (1993, 1994), Allenby and Rossi (1999), and references therein], classical and Bayesian flexible estimators of error component distributions [Horowitz and Markatou (1996), Chamberlain and Hirano (1999), Geweke and Keane (2000)], other nonparametric and semiparametric panel data models [Baltagi, Hidalgo and Li (1996), Li and Stengos (1996), Li and Hsiao (1998) and Chen, Heckman and Vytlacil (1998)], and models from time series of independent cross-sections [Deaton (1985), Moffitt (1993), Collado (1997)]. Some of these topics as well as comprehensive reviews of the panel data literature are covered in the text books by Hsiao (1986) and Baltagi (1995).

2. Linear models with predetermined variables: identification

In this section we discuss the identification of linear models with predetermined variables in two different contexts. In Section 2.1 the interest is to identify structural parameters in models in which explanatory variables are correlated with a time-invariant individual effect, but they are either strictly exogenous or predetermined relative to the time-varying errors. The second context, discussed in Section 2.2, is the time series analysis of error component models with autoregressive errors under various auxiliary assumptions. Section 2.3 discusses the use of stationarity restrictions in regression models, and Section 2.4 considers the identification of models with multiplicative or multiple individual effects.

2.1. Strict exogeneity, predeterminedness, and unobserved heterogeneity

We begin with a discussion of the implications of strict exogeneity for identification of regression parameters controlling for unobserved heterogeneity, with the objective of comparing this situation with that where the regressors are only predetermined variables.

Static regression with a strictly exogenous variable. Let us consider a linear regression for panel data including a fixed effect η_i and a time effect δ_t with N individuals observed T time periods, where T is small and N is large:

$$y_{it} = \beta x_{it} + \delta_t + \eta_i + \upsilon_{it} \quad (i = 1, \ldots, N;\ t = 1, \ldots, T). \tag{1}$$

We assume that $(y_{i1} \cdots y_{iT}, x_{i1} \cdots x_{iT}, \eta_i)$ is an iid random vector with finite second-order moments, while β and the time effects are treated as unknown parameters. The variable x_{it} is said to be strictly exogenous in this model if it is uncorrelated with past, present and future values of the disturbance υ_{it}:

$$E^*(\upsilon_{it}|x_i^T) = 0 \quad (t = 1, \ldots, T), \tag{2}$$

where E^* denotes a linear projection, and we use the superscript notation $z_i^t = (z_{i1}, \ldots, z_{it})'$. First-differencing the conditions we obtain

$$E^*(\upsilon_{it} - \upsilon_{i(t-1)}|x_i^T) = 0 \quad (t = 2, \ldots, T). \tag{3}$$

Since in the absence of any knowledge about η_i the condition $E^*(\upsilon_{i1}|x_i^T) = 0$ is not informative about β, the restrictions in first-differences are equivalent to those in levels. Therefore, for fixed T the problem of cross-sectional identification of β is simply that of a multivariate regression in first differences subject to cross-equation restrictions, and β is identifiable with $T \geqslant 2$.

Specifically, letting $E^*(\eta_i|x_i^T) = \lambda_0 + \lambda' x_i^T$, the model can be written as

$$y_{it} = \pi_{0t} + \beta x_{it} + \lambda' x_i^T + \varepsilon_{it} \text{ with } E^*(\varepsilon_{it}|x_i^T) = 0 \quad (t = 1, \ldots, T). \tag{4}$$

where $\pi_{0t} = \lambda_0 + \delta_t$. This T equation system is equivalent to

$$y_{i1} = \pi_{01} + \beta x_{i1} + \lambda' x_i^T + \varepsilon_{i1} \qquad E^*(\varepsilon_{i1}|x_i^T) = 0, \tag{5}$$

$$\Delta y_{it} = \Delta\delta_t + \beta\Delta x_{it} + \Delta\varepsilon_{it} \qquad E^*(\Delta\varepsilon_{it}|x_i^T) = 0 \quad (t = 2, \ldots, T). \tag{6}$$

In the absence of restrictions in λ Equation (5) is uninformative about β, and as a consequence asking under which conditions β is identified in Equation (4) is equivalent to asking under which conditions β is identified in Equation (6)[1].

[1] Lack of dependence between υ_{it} and x_i^T could also be expressed in terms of conditional independence in mean $E(\upsilon_{it}|x_i^T) = 0$ $(t = 1, \ldots, T)$. In the absence of any knowledge about η_i this is equivalent to the $(T - 1)$ conditional moment restrictions $E(\upsilon_{it} - \upsilon_{i(t-1)}|x_i^T) = 0$ $(t = 2, \ldots, T)$ which do not depend on η_i [Chamberlain (1992a)]. In the presentation for linear models, however, the use of linear projections affords a straightforward discussion of identification, and in the context of estimation it allows us to abstract from issues relating to optimal instruments and semiparametric asymptotic efficiency.

Partial adjustment with a strictly exogenous variable. In an alternative model, the effect of a strictly exogenous x on y could be specified as a partial adjustment equation:

$$y_{it} = \alpha y_{i(t-1)} + \beta_0 x_{it} + \beta_1 x_{i(t-1)} + \delta_t + \eta_i + v_{it} \quad (i = 1, \dots, N;\ t = 2, \dots, T) \tag{7}$$

together with

$$E^*(v_{it}|x_i^T) = 0 \quad (t = 2, \dots, T). \tag{8}$$

Note that assumption (8) does not restrict the serial correlation of v, so that lagged y is an endogenous explanatory variable. In the equation in levels, $y_{i(t-1)}$ will be correlated with η_i by construction and may also be correlated with past, present and future values of the errors v_{it} since they may be autocorrelated in an unspecified way. Likewise, the system in first differences is free from fixed effects and satifies $E^*(\Delta v_{it}|x_i^T) = 0$ ($t = 3, \dots, T$), but $\Delta y_{i(t-1)}$ may still be correlated with Δv_{is} for all s.

Subject to a standard rank condition, α, β_0, β_1 and the time effects will be identified with $T \geqslant 3$. With $T = 3$ they are just identified since there are five orthogonality conditions and five unknown parameters:

$$\begin{aligned} E\left[\begin{pmatrix} 1 \\ x_{i1} \\ x_{i2} \\ x_{i3} \end{pmatrix} (\Delta y_{i3} - \alpha \Delta y_{i2} - \beta_0 \Delta x_{i3} - \beta_1 \Delta x_{i2} - \Delta\delta_3)\right] &= 0. \\ E\,(y_{i2} - \alpha y_{i1} - \beta_0 x_{i2} - \beta_1 x_{i1} - \delta_2) &= 0. \end{aligned} \tag{9}$$

This simple example illustrates the potential for cross-sectional identification under strict exogeneity. In effect, strict exogeneity of x permits the identification of the dynamic effect of x on y and of lagged y on current y, in the presence of a fixed effect and shocks that can be arbitrarily persistent over time [cf. Bhargava and Sargan (1983), Chamberlain (1982a, 1984), Arellano (1990)].

A related situation of economic interest arises in testing life-cycle models of consumption or labor supply with habits [e.g., Bover (1991), or Becker, Grossman and Murphy (1994)]. In these models the coefficient on the lagged dependent variable is a parameter of central interest as it is intended to measure the extent of habits. However, in the absence of an exogenous instrumental variable such a coefficient would not be identified, since the effect of genuine habits could not be separated from serial correlation in the unobservables.

As an illustration, let us consider the empirical model of cigarette consumption by Becker, Grossman and Murphy (1994) for US state panel data. Their empirical analysis is based on the following equation:

$$c_{it} = \theta c_{i(t-1)} + \beta\theta c_{i(t+1)} + \gamma p_{it} + \eta_i + \delta_t + v_{i(t+1)}, \tag{10}$$

where c_{it} and p_{it} denote, respectively, annual per capita cigarette consumption in packs by state and average cigarette price per pack. Becker et al. are interested in testing whether smoking is addictive by considering the response of cigarette consumption to a change in cigarette prices.

The rationale for Equation (10) is provided by a model of addictive behavior in which utility in period t depends on cigarette consumption in t and in $t - 1$. Under perfect certainty and quadratic utility, the equation can be obtained from the first-order conditions of utility maximization. The degree of addiction is measured by θ, which will be positive if smoking is addictive. The current price coefficient γ should be negative by concavity of the utility, and β denotes the discount factor. With certainty, the marginal utility of wealth is constant over time but not cross-sectionally. The state specific intercept η_i is meant to capture such variation[2]. Finally, the δ_t's represent aggregate shocks, possibly correlated with prices, which are treated as period specific parameters.

The errors $\upsilon_{i(t+1)}$ capture unobserved life-cycle utility shifters, which are likely to be serially correlated. Therefore, even in the absence of addiction ($\theta = 0$) and serial correlation in prices, we would expect c_{it} to be autocorrelated, and in particular to find a non-zero effect of $c_{i(t-1)}$ in a linear regression of c_{it} on $c_{i(t-1)}$, $c_{i(t+1)}$ and p_{it}. Current consumption depends on prices in all periods through the effects of past and future consumption, but it is independent of past and future prices when $c_{i(t-1)}$ and $c_{i(t+1)}$ are held fixed. Thus, Becker et al.'s strategy is to identify θ, β, and γ from the assumption that prices are strictly exogenous relative to the unobserved utility shift variables. The required exogenous variation in prices comes from the variation in cigarette tax rates across states and time, and agents are assumed to be able to anticipate future prices without error.

Partial adjustment with a predetermined variable. The assumption that current values of x are not influenced by past values of y and υ is often unrealistic. We shall say that x is predetermined in a model like Equation (7) if

$$E^*(\upsilon_{it}|x_i^t, y_i^{t-1}) = 0 \quad (t = 2, \ldots, T). \tag{11}$$

That is, current shocks are uncorrelated with past values of y and with current and past values of x, but feedback effects from lagged dependent variables (or lagged errors) to current and future values of the explanatory variable are not ruled out.

Note that, in contrast with Equation (8), assumption (11) does restrict the serial correlation of υ. Specifically, it implies that the errors in first differences exhibit first-order autocorrelation but are uncorrelated at all other lags:

$$E(\Delta\upsilon_{it}\Delta\upsilon_{i(t-j)}) = 0 \quad j > 1.$$

Examples of this situation include Euler equations for household consumption [Zeldes (1989), Runkle (1991), Keane and Runkle (1992)], or for company investment

[2] According to the theory γ would also be state specific, since it is a function of the marginal utility of wealth. Thus the model with constant price coefficient must be viewed as an approximate model.

[Bond and Meghir (1994)], in which variables in the agents' information sets are uncorrelated with current and future idiosyncratic shocks but not with past shocks, together with the assumption that the empirical model's errors are given by such shocks.

Another example is the effect of children on female labour force participation decisions. In this context, assuming that children are strictly exogenous is much stronger than the assumption of predeterminedness, since it would require us to maintain that labour supply plans have no effect on fertility decisions at any point in the life cycle [Browning (1992, p. 1462)].

The implication of Equation (11) for errors in first differences is that

$$E^{*}(v_{it} - v_{i(t-1)} | x_i^{t-1}, y_i^{t-2}) = 0 \quad (t = 3, \ldots, T). \tag{12}$$

As before, these restrictions are equivalent to those in levels since in the absence of any knowledge about η_i the levels are not informative about the parameters[3]. Subject to a rank condition, α, β_0, β_1 and the time effects will be identified with $T \geqslant 3$. With $T = 3$ they are just identified from the five orthogonality conditions:

$$\begin{aligned} E\left[\begin{pmatrix} 1 \\ y_{i1} \\ x_{i1} \\ x_{i2} \end{pmatrix} (\Delta y_{i3} - \alpha \Delta y_{i2} - \beta_0 \Delta x_{i3} - \beta_1 \Delta x_{i2} - \Delta \delta_3)\right] &= 0, \\ E\,(y_{i2} - \alpha y_{i1} - \beta_0 x_{i2} - \beta_1 x_{i1} - \delta_2) &= 0. \end{aligned} \tag{13}$$

It is of some interest to compare the situation in Equation (13) with that in Equation (9). The two models are not nested since they only have four moment restrictions in common, which in this example are not sufficient to identify the five parameters. The model with a strictly exogenous x would become a special case of the model with a predetermined x, only if in the former serial correlation were ruled out. That is, if Equation (8) were replaced with:

$$E^{*}(v_{it} | x_i^{T}, y_i^{t-1}) = 0 \quad (t = 2, \ldots, T). \tag{14}$$

However, unlike in the predetermined case, lack of *arbitrary* serial correlation is not an identification condition for the model with strict exogeneity.

In the predetermined case it is still possible to accommodate special forms of serial correlation. For example, with $T = 4$ the parameters in the dynamic model are just identified with $E(\Delta v_{it} \Delta v_{i(t-j)}) = 0$ for $j > 2$, which is consistent with a first-order

[3] Orthogonality conditions of this type have been considered by Anderson and Hsiao (1981, 1982), Griliches and Hausman (1986), Holtz-Eakin, Newey and Rosen (1988), and Arellano and Bond (1991) amongst others.

moving average process for υ. This is so because in such case there are still three valid orthogonality restrictions: $E(y_{i1}\Delta\upsilon_{i4}) = 0$, $E(x_{i1}\Delta\upsilon_{i4}) = 0$, and $E(x_{i2}\Delta\upsilon_{i4}) = 0$.

Uncorrelated errors arise as the result of theoretical predictions in a number of environments (e.g., innovations in rational expectation models). However, even in the absence of specific restrictions from theory, the nature of shocks in econometric models is often less at odds with assumptions of no or limited autocorrelation than with the absence of feedback in the explanatory variable processes[4].

In the previous discussion we considered models for which the strict exogeneity property was unaffected by serial correlation, and models with feedback from lagged y or υ to current values of x, but other situations are possible. For example, it may be the case that the strict exogeneity condition (2) for model (1) is only satisfied as long as errors are unpredictable. An illustration is the agricultural Cobb–Douglas production function discussed by Chamberlain (1984), where y is log output, x is log labor, η is soil quality, and υ is rainfall. If η is known to farmers and they choose x to maximize expected profits, x will be correlated with η, but uncorrelated with υ at all lags and leads provided υ is unpredictable from past rainfall. If rainfall in t is predictable from rainfall in $t-1$, labour demand in t will in general depend on $\upsilon_{i(t-1)}$ [Chamberlain (1984, pp. 1258–1259)].

Another situation of interest is a case where the model is (1) or (7) and we only condition on x_i^t. That is, instead of Equation (11) we have

$$E^*(\upsilon_{it} \mid x_i^t) = 0. \tag{15}$$

In this case serial correlation is not ruled out, and the partial adjustment model is identifiable with $T \geqslant 4$, but Equation (15) rules out unspecified feedback from lagged y to current x. As an example, suppose that $\upsilon_{it} = \zeta_{it} + \varepsilon_{it}$ is an Euler equation's error given by the sum of a serially correlated preference shifter ζ_{it} and a white noise expectation error ε_{it}. The υ's will be serially correlated and correlated with lagged consumption variables y but not with lagged price variables x. Another example is an equation $y_{it}^* = \beta x_{it} + \eta_i + \upsilon_{it}^*$ where υ_{it}^* is white noise and x_{it} depends on $y_{i(t-1)}^*$, but y_{it}^* is measured with an autocorrelated error independent of x and y^* at all lags and leads.

Implications of uncorrelated effects. So far, we have assumed that all the observable variables are correlated with the fixed effect. If a strictly exogenous x were known to be uncorrelated with η, the parameter β in the static regression (1) would be identified from a single cross-section ($T = 1$). However, in the dynamic regression the lagged dependent variable would still be correlated with the effects by construction, so knowledge of lack of correlation between x and η would add T orthogonality conditions to the ones discussed above, but the parameters would still be identified

[4] As an example, see related discussions on the specification of shocks in Q investment equations by Hayashi and Inoue (1991), and Blundell, Bond, Devereux and Schiantarelli (1992).

only when $T \geqslant 3$[5]. The moment conditions for the partial adjustment model with strictly exogenous x and uncorrelated effects can be written as

$$E\left[\begin{pmatrix} 1 \\ x_i^T \end{pmatrix}(y_{it} - \alpha y_{i(t-1)} - \beta_0 x_{it} - \beta_1 x_{i(t-1)} - \delta_t)\right] = 0 \quad (t = 2, \ldots, T). \tag{16}$$

A predetermined x could also be known to be uncorrelated with the fixed effects if feedback occurred from lagged errors but not from lagged y. To illustrate this point suppose that the process for x is

$$x_{it} = \rho x_{i(t-1)} + \gamma v_{i(t-1)} + \phi\eta_i + \varepsilon_{it}, \tag{17}$$

where ε_{it}, v_{is} and η_i are mutually uncorrelated for all t and s. In this example x is uncorrelated with η when $\phi = 0$. However, if $v_{i(t-1)}$ were replaced by $y_{i(t-1)}$ in Equation (17), x and η will be correlated in general even with $\phi = 0$. Knowledge of lack of correlation between a predetermined x and η would also add T orthogonality restrictions to the ones discussed above for such a case. The moment conditions for the partial adjustment model with a predetermined x uncorrelated with the effects can be written as

$$\begin{aligned} E\left[\begin{pmatrix} 1 \\ x_i^t \end{pmatrix}\left(y_{it} - \alpha y_{i(t-1)} - \beta_0 x_{it} - \beta_1 x_{i(t-1)} - \delta_t\right)\right] &= 0 \quad (t = 2, \ldots, T), \\ E\left[y_i^{t-2}\left(\Delta y_{it} - \alpha\Delta y_{i(t-1)} - \beta_0\Delta x_{it} - \beta_1\Delta x_{i(t-1)} - \Delta\delta_t\right)\right] &= 0 \quad (t = 3, \ldots, T). \end{aligned} \tag{18}$$

Again, the parameters in this case would only be identified when $T \geqslant 3$.

Relationship with statistical definitions. To conclude this discussion, it may be useful to relate our usage of strict exogeneity to statistical definitions. A (linear projection based) *statistical* definition of strict exogeneity conditional on a fixed effect would state that x is strictly exogenous relative to y given η if

$$E^*(y_{it}|x_i^T, \eta_i) = E^*(y_{it}|x_i^t, \eta_i). \tag{19}$$

This is equivalent to the statement that y does not Granger-cause x given η in the sense that

$$E^*(x_{i(t+1)}|x_i^t, y_i^t, \eta_i) = E^*(x_{i(t+1)}|x_i^t, \eta_i). \tag{20}$$

Namely, letting $x_i^{(t+1)T} = (x_{i(t+1)}, \ldots, x_{iT})'$ if we have

$$E^*(y_{it}|x_i^T, \eta_i) = \beta_t' x_i^t + \delta_t' x_i^{(t+1)T} + \gamma_t\eta_i \tag{21}$$

and

$$E^*(x_{i(t+1)}|x_i^t, y_i^t, \eta_i) = \psi_t' x_i^t + \phi_t' y_i^t + \varsigma_t\eta_i, \tag{22}$$

it turns out that the restrictions $\delta_t = 0$ and $\phi_t = 0$ are equivalent. This result generalized the well-known equivalence between strict exogeneity [Sims (1972)] and Granger's

[5] Models with strictly exogenous variables uncorrelated with the effects were considered by Hausman and Taylor (1981), Bhargava and Sargan (1983), Amemiya and MaCurdy (1986), Breusch, Mizon and Schmidt (1989), Arellano (1993), and Arellano and Bover (1995).

non-causality [Granger (1969)][6]. It was due to Chamberlain (1984), and motivated the analysis in Holtz-Eakin, Newey and Rosen (1988), which was aimed at testing such a property.

Here, however, we are using strict exogeneity relative to the errors of an econometric model. Strict exogeneity itself, or the lack of it, may be a property of the model suggested by theory. We used some simple models as illustrations, in the understanding that the discussion would also apply to models that may include other features like individual effects uncorrelated with errors, endogenous explanatory variables, autocorrelation, or constraints in the parameters. Thus, in general strict exogeneity relative to a model may or may not be testable, but if so we shall usually be able to test it only in conjunction with other features of the model. In contrast with the econometric concept, a statistical definition of strict exogeneity is model free, but whether it is satisfied or not, may not necessarily be of relevance for the econometric model of interest[7].

As an illustration, let us consider a simple permanent-income model. The observables are non-durable expenditures c_{it}, current income w_{it}, and housing expenditure x_{it}. The unobservables are permanent (w_{it}^p) and transitory (ε_{it}) income, and measurement errors in non-durable (ξ_{it}) and housing (ς_{it}) expenditures. The expenditure variables are assumed to depend on permanent income only, and the unobservables are mutually independent but can be serially correlated. With these assumptions we have

$$w_{it} = w_{it}^p + \varepsilon_{it}, \tag{23}$$

$$c_{it} = \beta w_{it}^p + \xi_{it}, \tag{24}$$

$$x_{it} = \gamma w_{it}^p + \varsigma_{it}. \tag{25}$$

Suppose that β is the parameter of interest. The relationship between c_{it} and w_{it} suggested by the theory is of the form

$$c_{it} = \beta w_{it} + \upsilon_{it}, \tag{26}$$

where $\upsilon_{it} = \xi_{it} - \beta\varepsilon_{it}$. Since w_{it} and υ_{it} are contemporaneously correlated, w_{it} is an endogenous explanatory variable in Equation (26). Moreover, since $E^*(\upsilon_{it}|x_i^T) = 0$, x_{it} is a strictly exogenous instrumental variable in Equation (26). At the same time, note

[6] If linear projections are replaced by conditional distributions, the equivalence does not hold and it turns out that the definition of Sims is weaker than Granger's definition. Conditional Granger non-causality is equivalent to the stronger Sims' condition given by $f(y_t|x^T, y^{t-1}) = f(y_t|x^t, y^{t-1})$ [Chamberlain (1982b)].

[7] Unlike the linear predictor definition, a conditional independence definition of strict exogeneity given an individual effect is not restrictive, in the sense that there always exists a random variable η such that the condition is satisfied [Chamberlain (1984)]. This lack of identification result implies that a conditional-independence test of strict exogeneity given an individual effect will necessarily be a joint test involving a (semi) parametric specification of the conditional distribution.

that in general linear predictors of x given its past can be improved by adding lagged values of c and/or w (unless permanent income is white noise). Thus, the statistical condition for Granger non-causality or strict exogeneity is not satisfied in this example. A similar discussion could be conducted for a version of the model including fixed effects.

2.2. Time series models with error components

The motivation in the previous discussion was the identification of regression responses not contaminated from heterogeneity biases. Another leading motivation for using panel data is the analysis of the time series properties of the observed data. Models of this kind were discussed by Lillard and Willis (1978), MaCurdy (1982), Hall and Mishkin (1982), Holtz-Eakin, Newey and Rosen (1988) and Abowd and Card (1989), amongst others.

An important consideration is distinguishing unobserved heterogeneity from genuine dynamics. For example, the exercises cited above are all concerned with the time series properties of individual earnings for different reasons, including the analysis of earnings mobility, testing the permanent income hypothesis, or estimating intertemporal labour supply elasticities. However, how much dependence is measured in the residuals of the earnings process depends crucially, not only on how much heterogeneity is allowed into the process, but also on the auxiliary assumptions made in the specification of the residual process, and assumptions about measurement errors.

One way of modelling dynamics is through moving average processes [e.g., Abowd and Card (1989)]. These processes limit persistence to a fixed number of periods, and imply linear moment restrictions in the autocovariance matrix of the data. Autoregressive processes, on the other hand, imply nonlinear covariance restrictions but provide instrumental-variable orthogonality conditions that are linear in the autoregressive coefficients. Moreover, they are well suited to analyze the implications for identification and inference of issues such as the stationarity of initial conditions, homoskedasticity, and (near) unit roots.

Another convenient feature of autoregressive processes is that they can be regarded as a special case of the regression models with predetermined variables discussed above. This makes it possible to consider both types of problems in a common framework, and facilitates the distinction between static responses with residual serial correlation and dynamic responses[8]. Finally, autoregressive models are more easily extended to limited-dependent-variable models.

In the next subsection we discuss the implications for identification of alternative assumptions concerning a first-order autoregressive process with individual effects in short panels.

[8] In general, linear conditional models can be represented as data covariance matrix structures, but typically they involve a larger parameter space including many nuisance parameters, which are absent from instrumental-variable orthogonality conditions.

2.2.1. The AR(1) process with fixed effects[9]

Let us consider a random sample of individual time series of size T, $\{y_i^T, i = 1, \ldots, N\}$, with second-order moment matrix $E(y_i^T y_i^{T\prime}) = \Omega = \{\omega_{ts}\}$. We assume that the joint distribution of y_i^T and the individual effect η_i satisfies

$$y_{it} = \alpha y_{i(t-1)} + \eta_i + v_{it} \quad (i = 1, \ldots, N; \quad t = 2, \ldots, T) \quad |\alpha| < 1, \tag{27}$$

$$E^*(v_{it}|y_i^{t-1}) = 0 \quad (t = 2, \ldots, T), \tag{A1}$$

where $E(\eta_i) = \gamma$, $E(v_{it}^2) = \sigma_t^2$, and $\text{Var}(\eta_i) = \sigma_\eta^2$. Notice that the assumption does not rule out correlation between η_i and v_{it}, nor the possibility of conditional heteroskedasticity, since $E(v_{it}^2|y_i^{t-1})$ need not coincide with σ_t^2. Equations (27) and (A1) can be seen as a specialization of Equations (7) and (11). Thus, following the discussion above, (A1) implies $(T-2)(T-1)/2$ linear moment restrictions of the form

$$E[y_i^{t-2}(\Delta y_{it} - \alpha \Delta y_{i(t-1)})] = 0. \tag{28}$$

These restrictions can also be represented as constraints on the elements of Ω. Multiplying Equation (27) by y_{is} for $s < t$, and taking expectations gives $\omega_{ts} = \alpha\omega_{(t-1)s} + c_s$, $(t = 2, \ldots, T;\ s = 1, \ldots, t-1)$, where $c_s = E(y_{is}\eta_i)$. This means that, given assumption A1, the $T(T+1)/2$ different elements of Ω can be written as functions of the $2T \times 1$ parameter vector $\theta = (\alpha, c_1, \ldots, c_{T-1}, \omega_{11}, \ldots, \omega_{TT})'$. Notice that with $T = 3$ the parameters (α, c_1, c_2) are just identified as functions of the elements of Ω:

$$\begin{aligned}
\alpha &= (\omega_{21} - \omega_{11})^{-1}(\omega_{31} - \omega_{21}) \\
c_1 &= \omega_{21} - \alpha\omega_{11} \\
c_2 &= \omega_{32} - \alpha\omega_{22}.
\end{aligned}$$

The model based on A1 is attractive because the identification of α, which measures persistence given unobserved heterogeneity, is based on minimal assumptions. However, we may be willing to impose additional structure if this conforms to a priori beliefs.

Lack of correlation between the effects and the errors. One possibility is to assume that the errors v_{it} are uncorrelated with the individual effect η_i given y_i^{t-1}. In a structural context, this will often be a reasonable assumption if, for example, the v_{it} are interpreted as innovations that are independent of variables in the agents' information

[9] This section follows a similar discussion by Alonso-Borrego and Arellano (1999).

set. In such case, even if η_i is not observable to the econometrician, being time-invariant it is likely to be known to the individual. This situation gives rise to the following assumption

$$E^*(v_{it} | y_i^{t-1}, \eta_i) = 0 \quad (t = 2, \ldots, T). \tag{A1$'$}$$

Note that in a short panel assumption A1$'$ is more restrictive than assumption A1. Nevertheless, lack of correlation between v_{it} and $\{y_{i(t-1)}, \ldots, y_{i(t-J)}\}$ implies lack of correlation between v_{it} and η_i in the limit as $J \to \infty$. This will be so as long as

$$\eta_i = \operatorname*{plim}_{J \to \infty} \frac{1}{J} \sum_{j=1}^{J} \left(y_{i(t-j)} - \alpha y_{i(t-j-1)}\right).$$

Thus, for a process that started at $-\infty$ we would have orthogonality between η_i and v_{it}, and any correlation between individual effects and shocks will tend to vanish as t increases.

When $T \geqslant 4$, assumption A1$'$ implies the following additional $T - 3$ quadratic moment restrictions that were considered by Ahn and Schmidt (1995):

$$E[(y_{it} - \alpha y_{i(t-1)})(\Delta y_{i(t-1)} - \alpha \Delta y_{i(t-2)})] = 0 \quad (t = 4, \ldots, T). \tag{29}$$

In effect, we can write $E[(y_{it} - \alpha y_{i(t-1)} - \eta_i)(\Delta y_{i(t-1)} - \alpha \Delta y_{i(t-2)})] = 0$ and since $E(\eta_i \Delta v_{i(t-1)}) = 0$ the result follows. Thus, Equation (29) also holds if $\mathrm{Cov}(\eta_i, v_{it})$ is constant over t.

An alternative representation of the restrictions in Equation (29) is in terms of a recursion of the coefficients c_t introduced above. Multiplying Equation (27) by η_i and taking expectations gives $c_t = \alpha c_{t-1} + \phi, (t = 2, \ldots, T)$, where $\phi = E(\eta_i^2) = \gamma^2 + \sigma_\eta^2$, so that $c_1, \ldots, c_T$ can be written in terms of c_1 and ϕ. This gives rise to a covariance structure in which Ω depends on the $(T + 3) \times 1$ parameter vector $\theta = (\alpha, \phi, c_1, \omega_{11}, \ldots, \omega_{TT})'$. Notice that with $T = 3$ assumption A1$'$ does not imply further restrictions in Ω, with the result that α remains just identified. One can solve for ϕ in terms of α, c_1 and c_2:

$$\phi = (\omega_{32} - \omega_{21}) - \alpha(\omega_{22} - \omega_{11}).$$

Time series homoskedasticity. If in addition to A1$'$ we assume that the marginal variance of v_{it} is constant for all periods:

$$E(v_{it}^2) = \sigma^2 \quad (t = 2, \ldots, T), \tag{A2}$$

it turns out that

$$\omega_{tt} = \alpha^2 \omega_{(t-1)(t-1)} + \phi + \sigma^2 + 2\alpha c_{t-1} \quad (t = 2, \ldots, T).$$

This gives rise to a covariance structure in which Ω depends on five free parameters: $\alpha, \phi, c_1, \omega_{11}, \sigma^2$. This is a model of some interest since it is one in which the initial conditions of the process are unrestricted (governed by the parameters ϕ and c_1), but the total number of free parameters does not increase with T.

Mean stationarity of initial conditions. Other forms of additional structure that can be imposed are mean or variance stationarity conditions. The following assumption, which requires that the process started in the distant past, is a particularly useful mean stationarity condition:

$$\mathrm{Cov}(y_{it} - y_{i(t-1)}, \eta_i) = 0 \quad (t = 2, \ldots, T). \tag{B1}$$

Relative to assumption A1, assumption B1 adds the following $(T - 2)$ moment restrictions on Ω:

$$E[(y_{it} - \alpha y_{i(t-1)})\Delta y_{i(t-1)}] = 0 \quad (t = 3, \ldots, T), \tag{30}$$

which were proposed by Arellano and Bover (1995). However, relative to assumption A1′, assumption B1 only adds one moment restriction which can be written as $E[(y_{i3} - \alpha y_{i2})\Delta y_{i2}] = 0$. In terms of the parameters c_t, the implication of assumption B1 is that $c_1 = \cdots = c_T$ if we move from assumption A1, or that $c_1 = \phi/(1 - \alpha)$ if we move from assumption A1′. This gives rise to a model in which Ω depends on the $(T + 2) \times 1$ parameter vector $\theta = (\alpha, \phi, \omega_{11}, \ldots, \omega_{TT})'$. Notice that with $T = 3$, α is overidentified under assumption B1. Now α will also satisfy

$$\alpha = (\omega_{22} - \omega_{21})^{-1}(\omega_{32} - \omega_{31}).$$

It is of some interest to note that the combination of assumptions A1 and B1 produces the same model as that of A1′ and B1. However, while A1′ implies orthogonality conditions that are quadratic in α, A1 or A1+B1 give rise to linear instrumental-variable conditions [Ahn and Schmidt (1995)]. While A1 implied the validity of lagged levels as instruments for equations in first-differences, B1 additionally implies the validity of lagged first-differences as instruments for equations in levels. The availability of instruments for levels equations may lead to the identification of the effect of observable components of η_i (i.e., time-invariant regressors), or to identifying unit roots, two points to which we shall return below.

The validity of assumption B1 depends on whether initial conditions at the start of the sample are representative of the steady state behaviour of the model or not. For example, for young workers or new firms initial conditions may be less related to steady state conditions than for older ones.

Full stationarity. By combining A1′ with the homoskedasticity and the mean stationarity assumptions, A2 and B1, we obtain a model whose only nonstationary feature is the variance of the initial observation, which would remain a free parameter. For such a model $\omega_{tt} = \alpha^2\omega_{(t-1)(t-1)} + \sigma^2 + \phi(1 + \alpha)/(1 - \alpha)$ $(t = 2, \ldots, T)$. A fully stationary specification results from making the additional assumption:

$$\omega_{11} = \frac{\phi}{(1 - \alpha)^2} + \frac{\sigma^2}{(1 - \alpha^2)}. \tag{B2}$$

This gives rise to a model in which Ω only depends on the three parameters α, ϕ, and σ^2. Nevertheless, identification still requires $T \geqslant 3$, despite the fact that with

$T = 2$, Ω has three different coefficients. To see this, note that in their relationship to α, ϕ, and σ^2 the equation for the second diagonal term is redundant:

$$\omega_{tt} = \sigma^2_{\eta *} + \sigma^2_{\ell} \quad (t = 1, 2), \quad \omega_{12} = \alpha(\omega_{11} - \sigma^2_{\eta *}) + \sigma^2_{\eta *},$$

where $\sigma^2_{\eta *} = \sigma^2_{\eta}/(1 - \alpha)^2$ and $\sigma^2_{\ell} = \sigma^2/(1 - \alpha^2)$. The intuition for this is that both η_i and $y_{i(t-1)}$ induce serial correlation on y_{it}, but their separate effects can only be distinguished if at least first and second order autocorrelations are observed.

Under full stationarity (assumptions A1, A2, B1, and B2) it can be shown that

$$\frac{E(\Delta y_{i(t+1)} \Delta y_{it})}{E[(\Delta y_{it})^2]} = -\frac{(1 - \alpha)}{2}.$$

This is a well-known expression for the bias of the least squares regression in first-differences under homoskedasticity, which can be expressed as the orthogonality conditions

$$E\{\Delta y_{it}[(2y_{i(t+1)} - y_{it} - y_{i(t-1)}) - \alpha \Delta y_{it}]\} = 0 \quad (t = 2, \dots, T - 1).$$

With $T = 3$ this implies that α would also satisfy

$$\alpha = (\omega_{22} + \omega_{11} - 2\omega_{21})^{-1}[2(\omega_{32} - \omega_{31}) + \omega_{11} - \omega_{22}].$$

2.2.2. Aggregate shocks

Under assumptions A1 or A1′, the errors v_{it} are idiosyncratic shocks that are assumed to have cross-sectional zero mean at each point in time. However, if v_{it} contains aggregate shocks that are common to all individuals its cross-sectional mean will not be zero in general. This suggests replacing A1 with the assumption

$$E^*(v_{it} | y_i^{t-1}) = \delta_t \quad (t = 2, \dots, T), \tag{31}$$

which leads to an extension of the basic specification in which an intercept is allowed to vary over time:

$$y_{it} = \delta_t + \alpha y_{i(t-1)} + \eta_i + v^{\dagger}_{it}, \tag{32}$$

where $v^{\dagger}_{it} = v_{it} - \delta_t$. We can now set $E(\eta_i) = 0$ without lack of generality, since a nonzero mean would be subsumed in δ_t. Again, formally Equation (32) is just a specialization of Equations (7) and (11).

With fixed T, this extension does not essentially alter the previous discussion since the realized values of the shocks δ_t can be treated as unknown period specific parameters. With $T = 3$, α, δ_2 and δ_3 are just identified from the three moment conditions[10],

$$E(y_{i2} - \delta_2 - \alpha y_{i1}) = 0, \tag{33}$$

$$E(y_{i3} - \delta_3 - \alpha y_{i2}) = 0, \tag{34}$$

$$E[y_{i1}(\Delta y_{i3} - \Delta\delta_3 - \alpha\Delta y_{i2})] = 0. \tag{35}$$

In the presence of aggregate shocks the mean stationarity condition in assumption B1 may still be satisfied, but it will be interpreted as an assumption of mean stationarity conditional upon an aggregate effect (which may or may not be stationary), since now $E(\Delta y_{it})$ is not constant over t. The orthogonality conditions in Equation (30) remain valid in this case with the addition of a time varying intercept. With $T = 3$, assumption B1 adds to Equations (33–35) the orthogonality condition:

$$E[\Delta y_{i2}(y_{i3} - \delta_3 - \alpha y_{i2})] = 0. \tag{36}$$

2.2.3. Identification and unit roots

If one is interested in the unit root hypothesis, the model needs to be specified under both stable and unit roots environments. We begin by considering model (27) under assumption A1 as the stable root specification. As for the unit root specification, it is natural to consider a random walk without drift. The model can be written as

$$y_{it} = \alpha y_{i(t-1)} + (1 - \alpha)\eta_i^* + v_{it}, \tag{37}$$

where η_i^* denotes the steady state mean of the process when $|\alpha| < 1$. Thus, when $\alpha = 1$ we have

$$y_{it} = y_{i(t-1)} + v_{it}, \tag{38}$$

so that heterogeneity only plays a role in the determination of the starting point of the process. Note that in this model the covariance matrix of (y_{i1}, η_i^*) is left unrestricted.

An alternative unit root specification would be a random walk with an individual specific drift given by η_i:

$$y_{it} = y_{i(t-1)} + \eta_i + v_{it}, \tag{39}$$

but this is a model with heterogeneous linear growth that would be more suited for comparisons with stationary models that include individual trends.

[10] Further discussion on models with time effects is contained in Crepon, Kramarz and Trognon (1997).

The main point to notice here is that in model (37) α is not identified from the moments derived from assumption A1 when $\alpha = 1$. This is so because in the unit root case the lagged level will be uncorrelated with the current innovation, so that $\mathrm{Cov}(y_{i(t-2)}, \Delta y_{i(t-1)}) = 0$. As a result, the rank condition will not be satisfied for the basic orthogonality conditions (28). In model (39) the rank condition is still satisfied since $\mathrm{Cov}(y_{i(t-2)}, \Delta y_{i(t-1)}) \neq 0$ due to the cross-sectional correlation induced by the heterogeneity in shifts.

As noted by Arellano and Bover (1995), this problem does not arise when we consider a stable root specification that in addition to assumption A1 satisfies the mean stationarity assumption B1. The reason is that when $\alpha = 1$ the moment conditions (30) remain valid and the rank condition is satisfied since $\mathrm{Cov}(\Delta y_{i(t-1)}, y_{i(t-1)}) \neq 0$.

2.2.4. The value of information with highly persistent data

The cross-sectional regression coefficient of y_{it} on $y_{i(t-1)}$, ρ_t, can be expressed as a function of the model's parameters. For example, under full stationarity it can be shown to be

$$\rho = \alpha + \frac{\mathrm{Cov}(\eta_i, y_{i(t-1)})}{\mathrm{Var}(y_{i(t-1)})} = \alpha + \frac{(1-\alpha)\lambda^2}{\lambda^2 + (1-\alpha)/(1+\alpha)} \geqslant \alpha \tag{40}$$

where $\lambda = \sigma_\eta/\sigma$. Often, empirically ρ is near unity. For example, with firm employment data, Alonso-Borrego and Arellano (1999) found $\rho = 0.995$, $\alpha = 0.8$, and $\lambda = 2$. Since for any $0 \leqslant \alpha \leqslant \rho$ there is a value of λ such that ρ equals a pre-specified value, in view of lack of identification of α from the basic moment conditions (28) when $\alpha = 1$, it is of interest to see how the information about α in these moment conditions changes as T and α change for values of ρ close to one.

For the orthogonality conditions (28) the inverse of the semiparametric information bound about α can be shown to be

$$\sigma_T^2 = \sigma^2 \left\{ \sum_{s=1}^{T-2} E(y_{is}^* y_i^{s\prime})[E(y_i^s y_i^{s\prime})]^{-1} E(y_i^s y_{is}^*) \right\}^{-1} \tag{41}$$

where the y_{is}^* are orthogonal deviations relative to $(y_{i1}, \dots, y_{i(T-1)})'$[11]. The expression σ_T^2 gives the lower bound on the asymptotic variance of any consistent estimator of α based exclusively on the moments (28) when the process generating the data is the fully stationary model [Chamberlain (1987)].

[11] That is, y_{is}^* is given by $y_{is}^* = c_s[y_{is} - (T-s-1)^{-1}(y_{i(s+1)} + \cdots + y_{i(T-1)})]$ $(s = 1, \dots, T-2)$, where $c_s^2 = (T-s-1)/(T-s)$ [cf., Arellano and Bover (1995), and discussion in the next section].

Table 1
Inverse information bound for α (σ_T) when $\rho = 0.99$

T	σ_T [a]					
	(0, 9.9)	(0.2, 7.2)	(0.5, 4.0)	(0.8, 1.4)	(0.9, 0.7)	(0.99, 0)
3	14.14	15.50	17.32	18.97	19.49	19.95
4	1.97	2.66	4.45	8.14	9.50	10.00
5	1.21	1.55	2.43	4.71	5.88	6.34
10	0.50	0.57	0.71	1.18	1.61	1.85
15	0.35	0.38	0.44	0.61	0.82	0.96
Asympt. [b]	0.26	0.25	0.22	0.16	0.11	0.04

[a] Values for different (α, λ) pairs such that $\rho = 0.99$.
[b] Asymptotic standard deviation at $T = 15$, $\sqrt{(1 - \alpha^2)/15}$.

In Table 1 we have calculated values of σ_T for various values of T and for different pairs (α, λ) such that $\rho = 0.99$[12]. Also, the bottom row shows the time series asymptotic standard deviation, evaluated at $T = 15$, for comparisons.

Table 1 shows that with $\rho = 0.99$ there is a very large difference in information between $T = 3$ and $T > 3$. Moreover, for given T there is less information on α the closer α is to ρ. Often, there will be little information on α with $T = 3$ and the usual values of N. Additional information may be acquired from using some of the assumptions discussed above. Particularly, large gains can be obtained from employing mean stationarity assumptions, as suggested from Monte Carlo simulations reported by Arellano and Bover (1995) and Blundell and Bond (1998).

In making inferences about α we look for estimators whose sampling distribution for large N can be approximated by $N(\alpha, \sigma_T^2/N)$. However, there may be substantial differences in the quality of the approximation for a given N, among different estimators with the same asymptotic distribution. We shall return to these issues in the section on estimation.

2.3. *Using stationarity restrictions*

Some of the lessons from the previous section on alternative restrictions in autoregressive models are also applicable to regression models with predetermined (or strictly exogenous) variables of the form:

$$y_{it} = \delta' w_{it} + \eta_i + v_{it}, \tag{42}$$

$$E^*(v_{it} | w_i^t) = 0,$$

[12] Under stationarity σ_T^2 depends on α, λ and T but is invariant to σ^2.

where, e.g., $w_{it} = (y_{i(t-1)}, x_{it})'$. As before, the basic moments are $E[w_i^{t-1}(\Delta y_{it} - \delta'\Delta w_{it})] = 0$. However, if $E^*(v_{it}|w_i^t, \eta_i) = 0$ holds, the parameter vector δ also satisfies the Ahn–Schmidt restrictions

$$E[(y_{it} - \delta' w_{it})(\Delta y_{i(t-1)} - \delta \Delta w_{i(t-1)})] = 0. \tag{43}$$

Moreover, if $\mathrm{Cov}(\Delta w_{it}, \eta_i) = 0$ the Arellano–Bover restrictions are satisfied, encompassing the previous ones[13]:

$$E[\Delta w_{it}(y_{it} - \delta' w_{it})] = 0. \tag{44}$$

Blundell and Bond (1999) use moment restrictions of this type in their empirical analysis of Cobb–Douglas production functions using company panel data. They find that the instruments available for the production function in first differences are not very informative, due to the fact that the series on firm sales, capital and employment are highly persistent. In contrast, the first-difference instruments for production function errors in levels appear to be both valid and informative.

Sometimes the effect of time-invariant explanatory variables is of interest, a parameter γ, say, in a model of the form

$$y_{it} = \delta' w_{it} + \gamma z_i + \eta_i + v_{it}.$$

However, γ cannot be identified from the basic moments because the time-invariant regressor z_i is absorbed by the individual effect. Thus, we could ask whether the addition of orthogonality conditions involving errors in levels such as Equations (43) or (44) may help to identify such parameters. Unfortunately, often it would be difficult to argue that $E(\eta_i \Delta w_{it}) = 0$ without at the same time assuming that $E(z_i \Delta w_{it}) = 0$, in which case changes in w_{it} would not help the identification of γ. An example in which the levels restrictions may be helpful is the following simple model for an evaluation study due to Chamberlain (1993).

An evaluation of training example. Suppose that y_{it}^0 denotes earnings in the absence of training, and that there is a common effect of training for all workers. Actual earnings y_{it} are observed for $t = 1, \ldots, s-1, s+1, \ldots, T$. Training occurs in period s $(1 < s < T)$, so that $y_{it} = y_{it}^0$ for $t = 1, \ldots, s-1$, and we wish to measure its effect on earnings in subsequent periods, denoted by $\beta_{s+1}, \ldots, \beta_T$:

$$y_{it} = y_{it}^0 + \beta_t d_i \quad (t = s+1, \ldots, T), \tag{45}$$

where d_i is a dummy variable that equals 1 in the event of training. Moreover, we assume

$$y_{it}^0 = \alpha y_{i(t-1)}^0 + \eta_i + v_{it}, \tag{46}$$

[13] Strictly exogenous variables that had constant correlation with the individual effects were first considered by Bhargava and Sargan (1983).

together with $E^*(v_{it}|y_i^{0(t-1)}) = 0$ and $\mathrm{Cov}(\Delta y_{it}^0, \eta_i) = 0$. We also assume that d_i depends on lagged earnings $y_{i1}, \ldots, y_{i(s-1)}$ and η_i, but conditionally on these variables it is randomly assigned. Then we have:

$$y_{i(s+1)} = \alpha^2 y_{i(s-1)} + \beta_{s+1} d_i + (1+\alpha)\eta_i + (v_{i(s+1)} + \alpha v_{is}),$$

$$y_{it} = \alpha y_{i(t-1)} + (\beta_t - \alpha\beta_{t-1}) d_i + \eta_i + v_{it} \quad (t = s+2, \ldots, T).$$

From our previous discussion, the model implies the following orthogonality conditions:

$$E[y_i^{t-2}(\Delta y_{it} - \alpha \Delta y_{i(t-1)})] = 0 \quad (t = 1, \ldots, s-1), \tag{47}$$

$$E\{y_i^{s-2}[y_{i(s+1)} - (1+\alpha+\alpha^2) y_{i(s-1)} + \alpha(1+\alpha) y_{i(s-2)} - \beta_{s+1} d_i]\} = 0, \tag{48}$$

$$\begin{aligned} E\Bigg\{ y_i^{s-1} \Bigg[& y_{i(s+2)} - \frac{(1+\alpha+\alpha^2)}{(1+\alpha)} y_{i(s+1)} + \frac{\alpha^2}{(1+\alpha)} y_{i(s-1)} \\ & - \left(\beta_{i(s+2)} - \frac{(1+\alpha+\alpha^2)}{(1+\alpha)} \beta_{i(s+1)} \right) d_i \Bigg] \Bigg\} = 0. \end{aligned} \tag{49}$$

$$E[y_i^{t-2}(\Delta y_{it} - \alpha \Delta y_{i(t-1)} + \Delta(\beta_t - \alpha\beta_{t-1}) d_i)] = 0 \quad (t = s+3, \ldots, T). \tag{50}$$

The additional orthogonality conditions implied by mean stationarity are:

$$E[\Delta y_{i(t-1)}(y_{it} - \alpha y_{i(t-1)})] = 0 \quad (t = 1, \ldots, s-1), \tag{51}$$

$$E[\Delta y_{i(s-1)}(y_{i(s+1)} - \alpha^2 y_{i(s-1)} - \beta_{s+1} d_i)] = 0, \tag{52}$$

$$E[\Delta y_{i(s-1)}(y_{it} - \alpha y_{i(t-1)} + (\beta_t - \alpha\beta_{t-1}) d_i)] = 0 \quad (t = s+2, \ldots, T). \tag{53}$$

We would expect $E(\Delta y_{i(s-1)} d_i) < 0$, since there is evidence of a dip in the pretraining earnings of participants [e.g., Ashenfelter and Card (1985)]. Thus, Equation (52) can be expected to be more informative about β_{s+1} than Equation (48). Moreover, identification of β_{s+1} from Equation (48) requires that $s \geqslant 4$, otherwise only changes in β_t would be identified from Equations (47–50). In contrast, note that identification of β_{s+1} from Equation (52) only requires $s \geqslant 3$.

2.4. Models with multiplicative effects

In the models we have considered so far, unobserved heterogeneity enters exclusively through an additive individual specific intercept, while the other coefficients are assumed to be homogeneous. Nevertheless, an alternative autoregressive process could, for example, specify a homogeneous intercept and heterogeneity in the autoregressive behaviour:

$$y_{it} = \gamma + (\alpha + \eta_i) y_{i(t-1)} + v_{it}.$$

This is a potentially useful model if one is interested in allowing for agent specific adjustment cost functions, as for example in labour demand models. If we assume $E(v_{it} | y_i^{t-1}) = 0$ and $y_{it} > 0$, the transformed model,

$$y_{it} y_{i(t-1)}^{-1} = \gamma\, y_{i(t-1)}^{-1} + \alpha + \eta_i + v_{it}^{+},$$

where $v_{it}^{+} = v_{it} y_{i(t-1)}^{-1}$, also has $E(v_{it}^{+} | y_i^{t-1}) = 0$. Thus, the average autoregressive coefficient α and the intercept γ can be determined in a way similar to the linear models from the moment conditions $E(\eta_i + v_{it}^{+}) = 0$ and $E(y_i^{t-2} \Delta v_{it}^{+}) = 0$. Note that in this case, due to the nonlinearity, the argument requires the use of conditional mean assumptions as opposed to linear projections.

Another example is an exponential regression of the form

$$E(y_{it} | x_i^t, y_i^{t-1}, \eta_i) = \exp(\beta x_{it} + \eta_i).$$

This case derives its motivation from the literature on Poisson models for count data. The exponential specification is chosen to ensure that the conditional mean is always non-negative. With count data a log-linear regression is not a feasible alternative since a fraction of the observations on y_{it} will be zeroes.

A third example is a model where individual effects are interacted with time effects given by

$$y_{it} = \beta x_{it} + \delta_t \eta_i + v_{it}.$$

A model of this type may arise in the specification of unrestricted linear projections as in Equations (21) and (22), or as a structural specification in which an aggregate shock δ_t is allowed to have individual-specific effects on y_{it} measured by η_i.

Clearly, in such multiplicative cases first-differencing does not eliminate the unobservable effects, but as in the heterogeneous autoregression above there are simple alternative transformations that can be used to construct orthogonality conditions.

A transformation for multiplicative models. Generalizing the previous specifications we have

$$f_t(w_i^T, \gamma) = g_t(w_i^t, \beta)\eta_i + v_{it}, \qquad E(v_{it} | w_i^t) = 0, \tag{54}$$

where $g_{it} = g_t(w_i^t, \beta)$ is a function of predetermined variables and unknown parameters such that $g_{it} > 0$ for all w_i^t and β, and $f_{it} = f_t(w_i^T, \gamma)$ depends on endogenous and predetermined variables, as well as possibly also on unknown parameters. Dividing by g_{it} and first differencing the resulting equation, we obtain

$$f_{i(t-1)} - (g_{it}^{-1} g_{i(t-1)}) f_{it} = \upsilon_{it}^{+}, \tag{55}$$

and

$$E(\upsilon_{it}^{+} | w_i^{t-1}) = 0.$$

where $\upsilon_{it}^{+} = \upsilon_{i(t-1)} - (g_{it}^{-1} g_{i(t-1)}) \upsilon_{it}$.

Any function of w_i^{t-1} will be uncorrelated with υ_{it}^{+} and therefore can be used as an instrument in the determination of the parameters β and γ. This kind of transformation has been suggested by Chamberlain (1992b) and Wooldridge (1997). Notice that its use does not require us to condition on η_i. However, it does require g_t to be a function of predetermined variables as opposed to endogenous variables.

Multiple individual effects. We turn to consider models with more than one heterogeneous coefficient. Multiplicative random effects models with strictly exogenous variables were considered by Chamberlain (1992a), who found the information bound for a model with a multivariate individual effect. Chamberlain (1993) considered the identification problems that arise in models with predetermined variables when the individual effect is a vector with two or more components, and showed lack of identification of α in a model of the form

$$y_{it} = \alpha y_{i(t-1)} + \beta_i x_{it} + \eta_i + \upsilon_{it}, \tag{56}$$

$$E(\upsilon_{it} | x_i^t, y_i^{t-1}) = 0 \quad (t = 2, \ldots, T). \tag{57}$$

As an illustration consider the case where x_{it} is a $0 - 1$ binary variable. Since $E(\eta_i | x_i^T, y_i^{T-1})$ is unrestricted, the only moments that are relevant for the identification of α are

$$E(\Delta y_{it} - \alpha \Delta y_{i(t-1)} | x_i^{t-1}, y_i^{t-2}) = E(\beta_i \Delta x_{it} | x_i^{t-1}, y_i^{t-2}) \quad (t = 3, \ldots, T).$$

Letting $w_i^t = (x_i^t, y_i^t)$, the previous expression is equivalent to the following two conditions:

$$\begin{aligned} E(\Delta y_{it} - \alpha \Delta y_{i(t-1)} | w_i^{t-2}, x_{i(t-1)} = 0) = E(\beta_i | w_i^{t-2}, x_{i(t-1)} = 0) \\ \times \Pr(x_{it} = 1 | w_i^{t-2}, x_{i(t-1)} = 0), \end{aligned} \tag{58}$$

$$\begin{aligned} E(\Delta y_{it} - \alpha \Delta y_{i(t-1)} | w_i^{t-2}, x_{i(t-1)} = 1) = -E(\beta_i | w_i^{t-2}, x_{i(t-1)} = 1) \\ \times \Pr(x_{it} = 0 | w_i^{t-2}, x_{i(t-1)} = 1). \end{aligned} \tag{59}$$

Clearly, if $E(\beta_i | w_i^{t-2}, x_{i(t-1)} = 0)$ and $E(\beta_i | w_i^{t-2}, x_{i(t-1)} = 1)$ are unrestricted, and T is fixed, the autoregressive parameter α cannot be identified from Equations (58) and (59).

Let us consider some departures from model (56–57) under which α would be potentially identifiable. Firstly, if x were a strictly exogenous variable, in the sense that we replaced Equation (57) with the assumption $E(v_{it}|x_i^T, y_i^{t-1}) = 0$, α could be identifiable since

$$E(\Delta y_{it} - \alpha \Delta y_{i(t-1)}|x_i^T, y_i^{t-2}, \Delta x_{it} = 0) = 0. \tag{60}$$

Secondly, if the intercept η were homogeneous, identification of α and η could result from

$$E(y_{it} - \eta - \alpha y_{i(t-1)}|w_i^{t-1}, x_{it} = 0) = 0. \tag{61}$$

The previous discussion illustrates the fragility of the identification of dynamic responses from short time series of heterogeneous cross-sectional populations.

If $x_{it} > 0$ in model (56–57), it may be useful to discuss the ability of transformation (55) to produce orthogonality conditions. In this regard, a crucial aspect of the previous case is that while x_{it} is predetermined in the equation in levels, it becomes endogenous in the equation in first differences, so that transformation (55) applied to the first-difference equation does not lead to conditional moment restrictions. The problem is that although $E(\Delta v_{it}|x_i^{t-1}, y_i^{t-2}) = 0$, in general $E[(\Delta x_{it})^{-1}\Delta v_{it}|x_i^{t-1}, y_i^{t-2}] \neq 0$.

The parameters α, $\beta = E(\beta_i)$, and $\gamma = E(\eta_i)$ could be identifiable if x were a strictly exogenous variable such that $E(v_{it}|x_i^T, y_i^{t-1}) = 0$ $(t = 2, \dots, T)$, for in this case the transformed error $v_{it}^+ = (\Delta x_{it})^{-1}\Delta v_{it}$ would satisfy $E[v_{it}^+|x_i^T, y_i^{t-2}] = 0$ and $E[\Delta v_{it}^+|x_i^T, y_i^{t-3}] = 0$. Therefore, the following moment conditions would hold:

$$E\left[\left(\frac{\Delta y_{it}}{\Delta x_{it}} - \frac{\Delta y_{i(t-1)}}{\Delta x_{i(t-1)}}\right) - \alpha\left(\frac{\Delta y_{i(t-1)}}{\Delta x_{it}} - \frac{\Delta y_{i(t-2)}}{\Delta x_{i(t-1)}}\right) \middle| x_i^T, y_i^{t-3}\right] = 0, \tag{62}$$

$$E\left(\frac{\Delta y_{it}}{\Delta x_{it}} - \alpha\frac{\Delta y_{i(t-1)}}{\Delta x_{it}} - \beta\right) = 0, \tag{63}$$

$$\begin{aligned} E\Bigg[&\left(\frac{\Delta(y_{it}/x_{it})}{\Delta(1/x_{it})} - \frac{\Delta(y_{i(t-1)}/x_{i(t-1)})}{\Delta(1/x_{i(t-1)})}\right) \\ &- \alpha\left(\frac{\Delta(y_{i(t-1)}/x_{it})}{\Delta(1/x_{it})} - \frac{\Delta(y_{i(t-2)}/x_{i(t-1)})}{\Delta(1/x_{i(t-1)})}\right) \Bigg| x_i^T, y_i^{t-3}\Bigg] = 0, \end{aligned} \tag{64}$$

$$E\left(\frac{\Delta(y_{it}/x_{it})}{\Delta(1/x_{it})} - \alpha\frac{\Delta(y_{i(t-1)}/x_{it})}{\Delta(1/x_{it})} - \gamma\right) = 0. \tag{65}$$

A similar result would be satisfied if x_{it} in Equation (56) were replaced by a predetermined regressor that remained predetermined in the equation in first differences like $x_{i(t-1)}$. The result is that transformation (55) could be sequentially applied to models with predetermined variables and multiple individual effects, and still produce orthogonality conditions, as long as T is sufficiently large, and the transformed model resulting from the last but one application of the transformation still has the general form (54) (i.e., no functions of endogenous variables are multiplied by individual specific parameters).

A heterogeneous AR(1) model. As another example, consider a heterogeneous AR(1) model for a 0 – 1 binary indicator y_{it}:

$$\begin{aligned} y_{it} &= \eta_i + \alpha_i y_{i(t-1)} + v_{it}, \\ E(v_{it} | y_i^{t-1}) &= 0, \end{aligned} \tag{66}$$

and let us examine the (lack of) identification of the expected autoregressive parameter $E(\alpha_i)$ and the expected intercept $E(\eta_i)$. With $T = 3$, the only moment that is relevant for the identification of $E(\alpha_i)$ is

$$E(\Delta y_{i3} | y_{i1}) = E(\alpha_i \Delta y_{i2} | y_{i1}),$$

which is equivalent to the following two conditions:

$$E(\Delta y_{i3} | y_{i1} = 0) = E(\alpha_i | y_{i1} = 0, y_{i2} = 1) \Pr(y_{i2} = 1 | y_{i1} = 0), \tag{67}$$

$$E(\Delta y_{i3} | y_{i1} = 1) = -E(\alpha_i | y_{i1} = 1, y_{i2} = 0) \Pr(y_{i2} = 0 | y_{i1} = 1). \tag{68}$$

Therefore, only $E(\alpha_i | y_{i1} = 0, y_{i2} = 1)$ and $E(\alpha_i | y_{i1} = 1, y_{i2} = 0)$ are identified. The expected value of α_i for those whose value of y does not change from period 1 to period 2 is not identified, and hence $E(\alpha_i)$ is not identified either.

Similarly, for $T > 3$ we have

$$\begin{aligned} E(\Delta y_{it} | y_i^{t-3}, y_{i(t-2)} = 0) &= E(\alpha_i | y_i^{t-3}, y_{i(t-2)} = 0,\ y_{i(t-1)} = 1) \\ &\quad \times \Pr(y_{i(t-1)} = 1 | y_i^{t-3}, y_{i(t-2)} = 0), \\ E(\Delta y_{it} | y_i^{t-3}, y_{i(t-2)} = 1) &= -E(\alpha_i | y_i^{t-3}, y_{i(t-2)} = 1,\ y_{i(t-1)} = 0) \\ &\quad \times \Pr(y_{i(t-1)} = 0 | y_i^{t-3}, y_{i(t-2)} = 1). \end{aligned}$$

Note that $E(\alpha_i | y_i^{t-3}, y_{i(t-2)} = j,\ y_{i(t-1)} = j)$ for $j = 0, 1$ is also identified provided $E(\alpha_i | y_i^{t-3}, y_{i(t-2)} = j)$ is identified on the basis of the first $T - 1$ observations. The conclusion is that all conditional expectations of α_i are identified except $E(\alpha_i | y_{i1} = \cdots = y_{i(T-1)} = 1)$ and $E(\alpha_i | y_{i1} = \cdots = y_{i(T-1)} = 0)$.

Concerning η_i, note that since $E(\eta_i | y_i^{T-1}) = E(y_i^T | y_i^{T-1}) - y_{i(T-1)} E(\alpha_i | y_i^{T-1})$, expectations of the form $E(\eta_i | y_i^{T-2}, y_{i(T-1)} = 0)$ are all identified. Moreover, $E(\eta_i | y_i^{T-2}, y_{i(T-1)} = 1)$ is identified provided $E(\alpha_i | y_i^{T-2}, y_{i(T-1)} = 1)$ is identified. Thus, all conditional expectations of η_i are identified except $E(\eta_i | y_{i1} = \cdots = y_{i(T-1)} = 1)$.

Note that if $\Pr(y_{i1} = \cdots = y_{i(T-1)} = j)$ for $j = 0, 1$ tends to zero as T increases, $E(\alpha_i)$ and $E(\eta_i)$ will be identified as $T \to \infty$, but they may be seriously underidentified for very small values of T.

3. Linear models with predetermined variables: estimation

3.1. GMM estimation

Consider a model for panel data with sequential moment restrictions given by

$$\begin{aligned} y_{it} &= x_{it}'\beta_o + u_{it} \quad (t = 1, \ldots, T;\ i = 1, \ldots, N), \\ u_{it} &= \eta_i + v_{it}, \qquad E^*(v_{it} \mid z_i^t) = 0 \end{aligned} \tag{69}$$

where x_{it} is a $k \times 1$ vector of possibly endogenous variables, z_{it} is a $p \times 1$ vector of instrumental variables, which may include current values of x_{it} and lagged values of y_{it} and x_{it}, and $z_i^t = (z_{i1}', \ldots, z_{it}')'$. Observations across individuals are assumed to be independent and identically distributed. Alternatively, we can write the system of T equations for individual i as

$$y_i = X_i\beta_o + u_i, \tag{70}$$

where $y_i = (y_{i1}, \ldots, y_{iT})'$, $X_i = (x_{i1}', \ldots, x_{iT}')'$, and $u_i = (u_{i1}, \ldots, u_{iT})'$.

We saw that this model implies instrumental-variable orthogonality restrictions for the model in first-differences. In fact, the restrictions can be expressed using any $(T-1) \times T$ upper-triangular transformation matrix K of rank $(T-1)$, such that $K\iota = 0$, where ι is a $T \times 1$ vector of ones. Note that the first-difference operator is an example. We then have

$$E(Z_i'K u_i) = 0, \tag{71}$$

where Z_i is a block-diagonal matrix whose tth block is given by $z_i^{t\prime}$. An optimal GMM estimator of β_o based on Equation (71) is given by

$$\widehat{\beta} = (M_{zx}' A M_{zx})^{-1} M_{zx}' A M_{zy}, \tag{72}$$

where $M_{zx} = \left(\sum_{i=1}^N Z_i'K X_i\right)$, $M_{zy} = \left(\sum_{i=1}^N Z_i'K y_i\right)$, and A is a consistent estimate of the inverse of $E(Z_i'K u_i u_i'K' Z_i)$ up to a scalar. Under "classical" errors (that is, under conditional homoskedasticity $E(v_{it}^2 \mid z_i^t) = \sigma^2$, and lack of autocorrelation $E(v_{it} v_{i(t+j)} \mid z_i^{t+j}) = 0$ for $j > 0$), a "one-step" choice of A is optimal:

$$A_{\mathrm{C}} = \left(\sum_{i=1}^N Z_i'K K'Z_i\right)^{-1}. \tag{73}$$

Alternatively, the standard "two-step" robust choice is

$$A_{\mathrm{R}} = \left(\sum_{i=1}^N Z_i' K \widetilde{u}_i \widetilde{u}_i' K' Z_i\right)^{-1}, \tag{74}$$

where $\widetilde{u}_i = y_i - X_i\widetilde{\beta}$ is a vector of residuals evaluated at some preliminary consistent estimate $\widetilde{\beta}$.

Given identification, $\widehat{\beta}$ is consistent and asymptotically normal as $N \to \infty$ for fixed T [Hansen (1982)]. In addition, for either choice of A, provided the conditions under which they are optimal choices are satisfied, the asymptotic variance of $\widehat{\beta}$ is

$$\operatorname{Var}(\widehat{\beta})_{\mathrm{R}} = \{E(X_i'K'Z_i)[E(Z_i'K u_i u_i'K'Z_i)]^{-1}E(Z_i'KX_i)\}^{-1}, \tag{75}$$

which is invariant to K. Under classical errors this becomes[14]

$$\operatorname{Var}(\widehat{\beta})_{\mathrm{C}} = \sigma^2\{E(X_i'K'Z_i)[E(Z_i'KK'Z_i)]^{-1}E(Z_i'KX_i)\}^{-1}.$$

Moreover, as shown by Arellano and Bover (1995), a GMM estimator of the form given in Equations (72) and (73) or (74), is invariant to the choice of K provided K satisfies the required conditions [see also Schmidt, Ahn and Wyhowski (1992)].

As in common with other GMM estimation problems, the minimized estimation criterion provides an asymptotic chi-squared test statistic of the overidentifying restrictions. A two-step Sargan test statistic is given by

$$S_{\mathrm{R}} = \left[\sum_{i=1}^{N}(y_i - X_i\widehat{\beta}_{\mathrm{R}})'K'Z_i\right] A_{\mathrm{R}} \left[\sum_{i=1}^{N} Z_i'K(y_i - X_i\widehat{\beta}_{\mathrm{R}})\right] \to \chi^2_{(q-k)}, \tag{76}$$

where $\widehat{\beta}_{\mathrm{R}}$ is the two-step GMM estimator[15].

Orthogonal deviations. An alternative transformation to first differencing, which is very useful in the context of models with predetermined variables, is forward orthogonal deviations:

$$u_{it}^* = c_t\left[u_{it} - \frac{1}{(T-t)}(u_{i(t+1)} + \cdots u_{iT})\right], \tag{77}$$

where $c_t^2 = (T-t)/(T-t+1)$ [Arellano and Bover (1995)]. That is, to each of the first $(T-1)$ observations we subtract the mean of the remaining future observations available in the sample. The weighting c_t is introduced to equalize the variances of the transformed errors. A closely related transformation was used by Hayashi and Sims (1983) for time series models.

Unlike first differencing, which introduces a moving average structure in the error term, orthogonal deviations preserve lack of correlation among the transformed errors if the original ones are not autocorrelated and have constant variance. Indeed,

[14] Under classical errors, additional moment restrictions would be available, with the result that a smaller asymptotic variance could be achieved. The expression above simply particularizes the asymptotic variance to a situation where additional properties occur in the population but are not used in estimation.

[15] Similarly, letting $\widehat{\sigma}^2$ and $\widehat{\beta}_{\mathrm{C}}$ be, respectively, a consistent estimate of σ^2 and the one-step estimator, the one-step Sargan statistic is given by $S_{\mathrm{C}} = \widehat{\sigma}^{-2}\left[\sum_{i=1}^{N}(y_i - X_i\widehat{\beta}_{\mathrm{C}})'K'Z_i\right] A_{\mathrm{C}} \left[\sum_{i=1}^{N} Z_i'K(y_i - X_i\widehat{\beta}_{\mathrm{C}})\right]$.

orthogonal deviations can be regarded as the result of doing first differences to eliminate fixed effects plus a GLS transformation to remove the serial correlation induced by differencing.

The choice of K that produces this transformation is the forward orthogonal deviations operator $A = \mathrm{diag}[(T-1)/T, \ldots, 1/2]^{1/2} A^{+}$, where

$$A^{+} = \begin{pmatrix} 1 & -(T-1)^{-1} & -(T-1)^{-1} & \cdots & -(T-1)^{-1} & -(T-1)^{-1} & -(T-1)^{-1} \\ 0 & 1 & -(T-2)^{-1} & \cdots & -(T-2)^{-1} & -(T-2)^{-1} & -(T-2)^{-1} \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -1/2 & -1/2 \\ 0 & 0 & 0 & \cdots & 0 & 1 & -1 \end{pmatrix}.$$

It can be verified by direct multiplication that $AA' = I_{(T-1)}$ and $A'A = I_T - \iota\iota'/T \equiv Q$, which is the within-group operator. Thus, the OLS regression of y_{it}^* on x_{it}^* will give the within-group estimator, which is the conventional estimator in static models with strictly exogenous variables. Finally, since $Q = K'(KK')^{-1}K$, also $A = (KK')^{-1/2}K$ for any upper-triangular K.

A useful computational feature of orthogonal deviations, specially so when T is not a very small number, is that one-step estimators can be obtained as a matrix-weighted average of cross-sectional IV estimators:

$$\widehat{\beta} = \left(\sum_{t=1}^{T-1} X_t^{*\prime} Z_t (Z_t' Z_t)^{-1} Z_t' X_t^* \right)^{-1} \sum_{t=1}^{T-1} X_t^{*\prime} Z_t (Z_t' Z_t)^{-1} Z_t' y_t^*, \tag{78}$$

where $X_t^* = (x_{1t}^{*\prime}, \ldots, x_{Nt}^{*\prime})'$, $y_t^* = (y_{1t}^*, \ldots, y_{Nt}^*)'$, and $Z_t = (z_i^{t\prime}, \ldots, z_N^{t\prime})'$.

An illustration: female labour force participation and fertility. We illustrate the previous issues with reference to an empirical relationship between female participation and fertility, discussing a simplified version of the results reported by Carrasco (1998) for a linear probability model[16].

A sample from PSID for 1986–1989 is used. The data consists of 1442 women aged 18–55 in 1986, that are either married or cohabiting. The left-hand side variable is a binary indicator of participation in year t. Fertility is also a dummy variable, which takes the value one if the age of the youngest child in $t+1$ is 1. The equation also includes an indicator of whether the woman has a child aged 2–6. The equations estimated in levels also include a constant, age, race, and education dummies (not reported).

In this sample it is observed that women with two children of the same sex have a significantly higher probability of having a third child. Thus, the sex of the first two children is used as an instrument for fertility, which is treated as an endogenous

[16] We thank Raquel Carrasco for allowing us to draw freely on her dataset and models.

Table 2
Linear probability models of female labour force participation[a,b] (N = 1442, 1986–1989)

Variable	OLS	2SLS[c]	WITHIN	GMM[d]	GMM[e]
Fertility	–0.15	–1.01	–0.06	–0.08	–0.13
	(8.2)	(2.1)	(3.8)	(2.8)	(2.2)
Kids 2–6	–0.08	–0.24	0.001	–0.005	–0.09
	(5.2)	(2.6)	(0.04)	(0.4)	(2.7)
Sargan test				48.0 (22)	18.0 (10)
$m1$	19.0	5.7	–10.0	–10.0	–10.0
$m2$	16.0	12.0	–1.7	–1.7	–1.6
Models including lagged participation					
Fertility	–0.09	–0.33	–0.06	–0.09	–0.14
	(5.2)	(1.3)	(3.7)	(3.1)	(2.2)
Kids 2–6	–0.02	–0.07	–0.000	–0.02	–0.10
	(2.1)	(1.3)	(0.00)	(1.1)	(3.5)
Lagged participation	0.63	0.61	0.03	0.36	0.29
	(42.0)	(30.0)	(1.7)	(8.3)	(6.3)
Sargan				51.0 (27)	25.0 (15)
$m1$	–7.0	–5.4	–13.0	–14.0	–13.0
$m2$	3.1	2.8	–1.3	1.5	1.2

[a] Heteroskedasticity robust t-ratios shown in parentheses.
[b] GMM IVs in bottom panel also include lags of participation up to $t-2$.
[c] External instrument: previous children of same sex.
[d] IVs: all lags and leads of "kids 2–6" and "same sex" variables (strictly exogenous).
[e] IVs: lags of "kids 2–6" and "same sex" up to $t-1$ (predetermined).

variable. The presence of a child aged 2–6 is the result of past fertility decisions, and so it should be treated as a predetermined variable [see Carrasco (1998) for a comprehensive discussion, and additional estimates of linear and nonlinear models].

Table 2 reports the results for two versions of the model with and without lagged participation as a regressor, using DPD [Arellano and Bond (1988)]. The last column presents GMM estimates in orthogonal deviations that treat fertility as endogenous, and the "kids 2–6" and "same sex" indicators as predetermined variables. The table also reports the results from other methods of estimation for comparisons.

There is a large gap between the OLS and 2SLS measured effects of fertility, possibly due to measurement errors. Both OLS and 2SLS neglect unobserved heterogeneity, despite evidence from the serial correlation statistics $m1$ and $m2$ of persistent positive autocorrelation in the residuals in levels. Note that we would expect the "same sex" instrumental variable to be correlated with the fixed effect. The reason is that it will be a predictor of preferences for children, given that the sample includes women with less than two children.

The within-groups estimator controls for unobserved heterogeneity, but in doing so we would expect it to introduce biases due to lack of strict exogeneity of the explanatory variables. The GMM estimates in column 4 deal with the endogeneity of fertility and control for fixed effects, but treat the "kids 2–6" and "same sex" variables as strictly exogenous. This results in a smaller effect of fertility on participation (in absolute value) than the one obtained in column 5 treating the variables as predetermined. The hypothesis of strict exogeneity of these two variables is rejected at the 5 percent level from the difference in the Sargan statistics in both panels. (Both GMM estimates are "one-step", but all test statistics reported are robust to heteroskedasticity.)

Finally, note that the $m1$ and $m2$ statistics (which are asymptotically distributed as a $N(0, 1)$ under the null of no autocorrelation) have been calculated from residuals in first differences for the within-groups and GMM estimates. So if the errors in levels were uncorrelated, we would expect $m1$ to be significant, but not $m2$, as is the case here [cf., Arellano and Bond (1991)].

Levels and differences estimators. The GMM estimator proposed by Arellano and Bover (1995) combined the basic moments (71) with $E(\Delta z_{it} u_{it}) = 0$ ($t = 2, \dots, T$). Using their notation, the full set of orthogonality conditions can be written in compact form as

$$E(Z_i^{+\prime} H u_i) = 0, \tag{79}$$

where Z_i^{+} is a block diagonal matrix with blocks Z_i as above, and $Z_{\ell i}$ = diag $(\Delta z'_{i2}, \dots, \Delta z'_{iT})$. H is the $2(T-1) \times T$ selection matrix $H = (K', I_o')'$, where $I_o = (0 \vdots I_{T-1})$. With these changes in notation, the form of the estimator is similar to that in Equation (72).

As before, a robust choice of A is provided by the inverse of an unrestricted estimate of the variance matrix of the moments $N^{-1} \sum_{i=1}^{N} Z_i^{+\prime} H \widetilde{u}_i \widetilde{u}_i' H' Z_i^{+}$. However, this can be a poor estimate of the population moments if N is not sufficiently large relative to T, which may have an adverse effect on the finite sample properties of the GMM estimator. Unfortunately, in this case an efficient one-step estimator under restrictive assumptions does not exist. Intuitively, since some of the instruments for the equations in levels are not valid for those in differences, and conversely, not all the covariance terms between the two sets of moments will be zero.

3.2. Efficient estimation under conditional mean independence

If lack of correlation between v_{it} and z_i^t is replaced by an assumption of conditional independence in mean $E(v_{it} \mid z_i^t) = 0$, the model implies additional orthogonality restrictions. This is so because v_{it} will be uncorrelated not only with the conditioning variables z_i^t but also with functions of them. Chamberlain (1992b) derived the semiparametric efficiency bound for this model. Hahn (1997) showed that a GMM estimator based on an increasing set of instruments as N tends to infinity would achieve the semiparametric efficiency bound. Hahn discussed the rate of growth of the number of instruments for the case of Fourier series and polynomial series.

Note that the asymptotic bound for the model based on $E(v_{it} \,|\, z_i^t) = 0$ will be in general different from that of $E(v_{it} \,|\, z_i^t, \eta_i) = 0$, whose implications for linear projections were discussed in the previous section.

Similarly, the bound for a version of the model with levels and differences restrictions based on conditional mean independence assumptions cannot be obtained either as an application of Chamberlain's results. The reason is that the addition of the level's conditions breaks the sequential moment structure of the problem.

Let us now consider the form of the information bound and the optimal instruments for model (69) together with the conditional mean assumption $E(v_{it} \,|\, z_i^t) = 0$. Since $E(\eta_i \,|\, z_i^T)$ is unrestricted, all the information about β is contained in $E(v_{it} - v_{i(t+1)} \,|\, z_i^t) = 0$ for $t = 1, \ldots, T-1$.

For a single period the information bound is $J_{0t} = E(d_{it} d_{it}' / \omega_{it})$ where $d_{it} = E(x_{it} - x_{i(t+1)} \,|\, z_i^t)$ and $\omega_{it} = E[(v_{it} - v_{i(t+1)})^2 \,|\, z_i^t]$ [cf., Chamberlain (1987)]. Thus, for a single period the optimal instrument is $m_{it} = d_{it} / \omega_{it}$, in the sense that under suitable regularity conditions the statistic

$$\widetilde{\beta}_{(t)} = \left(\sum_{i=1}^{N} m_{it} \Delta x_{i(t+1)}' \right)^{-1} \left(\sum_{i=1}^{N} m_{it} \Delta y_{i(t+1)} \right),$$

satisfies $\sqrt{N}(\widetilde{\beta}_{(t)} - \beta) \xrightarrow{d} N(0, J_{0t}^{-1})$. If the errors were conditionally serially uncorrelated, the total information would be the sum of the information bounds for each period. So Chamberlain (1992b) proposed the following recursive forward transformation of the first-differenced errors:

$$\begin{aligned} \tilde{v}_{i(T-1)} &= v_{i(T-1)} - v_{iT}, \\ \tilde{v}_{it} &= (v_{it} - v_{i(t+1)}) \\ &\quad - \frac{E[(v_{it} - v_{i(t+1)})\tilde{v}_{i(t+1)} \,|\, z_i^{t+1}]}{E(\tilde{v}_{i(t+1)}^2 \,|\, z_i^{t+1})} \tilde{v}_{i(t+1)} \\ &\quad - \frac{E[(v_{it} - v_{i(t+1)})\tilde{v}_{i(t+2)} \,|\, z_i^{t+2}]}{E(\tilde{v}_{i(t+2)}^2 \,|\, z_i^{t+2})} \tilde{v}_{i(t+2)} \\ &\quad - \cdots \\ &\quad - \frac{E[(v_{it} - v_{i(t+1)})\tilde{v}_{i(T-1)} \,|\, z_i^{T-1}]}{E(\tilde{v}_{i(T-1)}^2 \,|\, z_i^{T-1})} \tilde{v}_{i(T-1)}, \end{aligned} \tag{80}$$

for $t = T-2, \ldots, 1$. The interest in this transformation is that it satisfies the same conditional moment restrictions as the original errors in first-differences, namely

$$E(\tilde{v}_{it} \,|\, z_i^t) = 0, \tag{81}$$

but additionally it satisfies by construction the lack of dependence requirement:

$$E(\tilde{\upsilon}_{it}\,\tilde{\upsilon}_{i\,(t+j)}\,|\,z_i^{t+j}) = 0 \text{ for } j = 1, \ldots, T-t-1. \tag{82}$$

Therefore, in terms of the transformed errors the information bound can be written as

$$J_0 = \sum_{t=1}^{T-1} E(\widetilde{d}_{it}\,\widetilde{d}'_{it}/\widetilde{\omega}_{it}), \tag{83}$$

where $\widetilde{d}_{it} = E(\tilde{x}_{it}\,|\,z_i^t)$ and $\widetilde{\omega}_{it} = E(\tilde{\upsilon}_{it}^2\,|\,z_i^t)$. The variables $\tilde{x}_{it}$ and $\tilde{y}_{it}$ denote the corresponding transformations to the first-differences of x_{it} and y_{it} such that $\tilde{\upsilon}_{it} = \tilde{y}_{it} - \tilde{x}'_{it}\beta$. Thus, the optimal instruments for all periods are $\widetilde{m}_{it} = \widetilde{d}_{it}/\widetilde{\omega}_{it}$, in the sense that under suitable regularity conditions the statistic

$$\widetilde{\beta} = \left(\sum_{i=1}^{N}\sum_{t=1}^{T-1} \widetilde{m}_{it}\,\tilde{x}'_{it}\right)^{-1} \left(\sum_{i=1}^{N}\sum_{t=1}^{T-1} \widetilde{m}_{it}\,\tilde{y}_{it}\right)$$

satisfies $\sqrt{N}(\widetilde{\beta} - \beta) \xrightarrow{d} N(0, J_0^{-1})$.

If the υ_{it}'s are conditionally homoskedastic and serially uncorrelated, so that $E(\upsilon_{it}^2\,|\,z_i^t) = \sigma^2$ and $E(\upsilon_{it}\upsilon_{i\,(t+j)}\,|\,z_i^{t+j}) = 0$ for $j > 0$, it can be easily verified that the $\tilde{\upsilon}_{it}$'s blow down to ordinary forward orthogonal deviations as defined in Equation (77):

$$\tilde{\upsilon}_{it} = \upsilon_{it} - \frac{1}{(T-t)}(\upsilon_{i\,(t+1)} + \cdots + \upsilon_{iT}) \equiv \frac{1}{c_t}\upsilon_{it}^* \text{ for } t = T-1, \ldots, 1.$$

In such case $\widetilde{m}_{it} = c_t\sigma^{-2}E(x_{it}^*\,|\,z_i^t)$ so that

$$\widetilde{\beta} = \left(\sum_{i=1}^{N}\sum_{t=1}^{T-1} E(x_{it}^*\,|\,z_i^t)x_{it}^{*\prime}\right)^{-1} \left(\sum_{i=1}^{N}\sum_{t=1}^{T-1} E(x_{it}^*\,|\,z_i^t)y_{it}^*\right), \tag{84}$$

and

$$J_0 = \frac{1}{\sigma^2}\sum_{t=1}^{T-1} E[E(x_{it}^*\,|\,z_i^t)E(x_{it}^{*\prime}\,|\,z_i^t)]. \tag{85}$$

If we further assume that the conditional expectations $E(x_{it}^*\,|\,z_i^t)$ are linear, then

$$J_0 = \frac{1}{\sigma^2}\sum_{t=1}^{T-1} E(x_{it}^* z_i^{t\prime})[E(z_i^t z_i^{t\prime})]^{-1}E(z_i^t x_{it}^{*\prime}), \tag{86}$$

which coincides with the inverse of the asymptotic covariance matrix of the simple IV estimator given in Equation (78) under the stated assumptions. Note that the assumptions of conditional homoskedasticity, lack of serial correlation, and linearity of $E(x_{it}^* \,|\, z_i^t)$ would imply further conditional moment restrictions that may lower the information bound for β. Here, we merely particularize the bound for β based on $E(v_{it} \,|\, z_i^t) = 0$ to the case where the additional restrictions happen to occur in the population but are not used in the calculation of the bound.

3.3. Finite sample properties of GMM and alternative estimators

For sufficiently large N, the sampling distribution of the GMM estimators discussed above can be approximated by a normal distribution. However, the quality of the approximation for a given sample size may vary greatly depending on the quality of the instruments used. Since the number of instruments increases with T, many overidentifying restrictions tend to be available even for moderate values of T, although the quality of these instruments is often poor.

Monte Carlo results on the finite sample properties of GMM estimators for panel data models with predetermined variables have been reported by Arellano and Bond (1991), Kiviet (1995), Ziliak (1997), Blundell and Bond (1998) and Alonso-Borrego and Arellano (1999), amongst others. A conclusion in common to these studies is that GMM estimators that use the full set of moments available for errors in first-differences can be severely biased, specially when the instruments are weak and the number of moments is large relative to the cross-sectional sample size.

From the literature on the finite sample properties of simultaneous equations estimators, we know that the effect of weak instruments on the distributions of 2SLS and LIML differs substantially, in spite of the fact that both estimators have the same asymptotic distribution. While LIML is approximately median unbiased, 2SLS is biased towards OLS, and in the case of lack of identification in the population it converges to a random variable with the OLS probability limit as its central value. In contrast, LIML has no moments, and as a result its distribution has thicker tails than that of 2SLS and a higher probability of outliers [cf., Phillips (1983)]. Anderson, Kunitomo and Sawa (1982) carried out numerical comparisons of the distributions of the two estimators, and concluded that LIML was to be strongly preferred to 2SLS, specially in cases with a large number of instruments.

LIML analogue estimators. It is thus of interest to consider LIML analogues for our models, and compare their finite sample properties with those of GMM estimators. Following Alonso-Borrego and Arellano (1999), a non-robust LIML analogue $\widehat{\beta}_{\mathrm{LIML1}}$ minimizes a criterion of the form

$$\ell_{\mathrm{C}}(\beta) = \frac{(y^* - X^*\beta)' M (y^* - X^*\beta)}{(y^* - X^*\beta)'(y^* - X^*\beta)}, \tag{87}$$

where starred variables denote orthogonal deviations, $y^* = (y_1^{*\prime}, \ldots, y_N^{*\prime})'$, $X^* = (X_1^{*\prime}, \ldots, X_N^{*\prime})'$, $Z = (Z_1', \ldots, Z_N')'$, and $M = Z(Z'Z)^{-1}Z'$. The resulting estimator is

$$\widehat{\beta}_{\mathrm{LIML1}} = (X^{*\prime} M X^* - \hat{\ell} X^{*\prime} X^*)^{-1} (X^{*\prime} M y^* - \hat{\ell} X^{*\prime} y^*), \tag{88}$$

where $\hat{\ell}$ is the minimum eigenvalue of the matrix $W^{*\prime}MW^{*}(W^{*\prime}W^{*})^{-1}$, and $W^{*} = (y^{*}, X^{*})$.

The estimator in Equation (88) is algebraically similar to an ordinary single-equation LIML estimator provided the model is in orthogonal deviations. This is so in spite of having a system of equations, due to the fact that the errors in orthogonal deviations of different equations are serially uncorrelated and homoskedastic under classical assumptions. However, the non-robust LIML analogue does not correspond to any meaningful maximum likelihood estimator (for example, it does not exploit the homoskedasticity restrictions). It is only a "LIML" estimator in the sense of the instrumental-variable interpretation given by Sargan (1958) to the original LIML estimator, and generalized to robust contexts by Hansen, Heaton and Yaron (1996).

The robust LIML analogue $\widehat{\beta}_{\mathrm{LIML2}}$, or continuously updated GMM estimator in the terminology of Hansen et al. (1996), minimizes a criterion of the form

$$\ell_{\mathrm{R}}(\beta) = (y^{*} - X^{*}\beta)'Z\left(\sum_{i=1}^{N} Z_i' u_i^{*}(\beta) u_i^{*}(\beta)' Z_i\right)^{-1} Z'(y^{*} - X^{*}\beta), \tag{89}$$

where $u_i^{*}(\beta) = y_i^{*} - X_i^{*}\beta$. Note that LIML2, unlike LIML1, does not solve a standard minimum eigenvalue problem, and requires the use of numerical optimization methods[17].

In contrast to GMM, the LIML estimators are invariant to normalization. Hillier (1990) showed that the alternative normalization rules adopted by LIML and 2SLS were at the root of their different sampling properties. He also showed that a symmetrically normalized 2SLS estimator had similar properties to those of LIML. Alonso-Borrego and Arellano (1999) considered symmetrically normalized GMM (SNM) estimators for panel data, and compared them with ordinary GMM and LIML analogues by mean of simulations. The main advantage of robust SNM over robust LIML is computational, since the former solves a minimum eigenvalue problem while the latter does not. It also avoids potential problems of non-convergence with LIML2, as reported by Alonso-Borrego and Arellano (1999).

The Monte Carlo results and the empirical illustrations for autoregressive models reported by Alonso-Borrego and Arellano (1999) showed that GMM estimates can exhibit large biases when the instruments are poor, while the symmetrically normalized estimators (LIML and SNM) remained essentially unbiased. However, LIML and SNM always had a larger interquartile range than GMM, although the differences were small except in the almost unidentified cases.

[17] Other one-step methods that achieve the same asymptotic efficiency as robust GMM or LIML estimators are the empirical likelihood [Back and Brown (1993), Qin and Lawless (1994) and Imbens (1997)] and exponential tilting estimators [Imbens, Spady and Johnson (1998)]. Nevertheless, little is known as yet on the relative merits of these estimators in panel data models, concerning computational aspects and their finite sample properties.

3.4. Approximating the distributions of GMM and LIML for AR(1) models when the number of moments is large

Within-groups estimators of autoregressive models, and more generally of models with predetermined variables, are known to be consistent as T tends to infinity, but are inconsistent for fixed T and large N [cf., Nickell (1981), Anderson and Hsiao (1981)]. On the other hand, the estimators reviewed above are consistent for fixed T but the number of orthogonality conditions increases with T. In panels in which the value of T is not negligible relative to N (such as the PSID household incomes panel in the US, or the balance sheet-based company panels that are available in many countries), the knowledge of the asymptotic behaviour of the estimators as both T and N tend to infinity may be useful in assessing alternative methods.

Alvarez and Arellano (1998) obtained the asymptotic properties of within-groups (WG), one-step GMM, and non-robust LIML for a first-order autoregressive model when both N and T tend to infinity. Hahn (1998) also obtained the asymptotic properties of WG under more general conditions. The main results can be summarized in the following proposition.

Proposition 1. *Let* $y_{it} = \alpha y_{i(t-1)} + \eta_i + v_{it}$, *with* $v_{it}|y_i^{t-1}, \eta_i \sim i.i.d.N(0, \sigma^2)$, $(t = 1, \ldots, T)$ *and* $y_{i0}|\eta_i \sim N[\eta_i/(1-\alpha), \sigma^2/(1-\alpha^2)]$. *Also let* $\eta_i \sim i.i.d.N(0, \sigma_\eta^2)$. *Then, as both N and T tend to infinity, provided* $T/N \to c$, $0 \leqslant c \leqslant 2$, *within-groups, GMM1, and LIML1 are consistent for* α. *Moreover,*

$$\sqrt{NT}\left[\widehat{\alpha}_{\text{GMM1}} - \left(\alpha - \frac{1}{N}(1+\alpha)\right)\right] \xrightarrow{d} N(0, 1-\alpha^2), \tag{90}$$

$$\sqrt{NT}\left[\widehat{\alpha}_{\text{LIML1}} - \left(\alpha - \frac{1}{(2N-T)}(1+\alpha)\right)\right] \xrightarrow{d} N(0, 1-\alpha^2). \tag{91}$$

Also, provided $N/T^3 \to 0$*:*

$$\sqrt{NT}\left[\widehat{\alpha}_{\text{WG}} - \left(\alpha - \frac{1}{T}(1+\alpha)\right)\right] \xrightarrow{d} N(0, 1-\alpha^2). \tag{92}$$

Proof: See Alvarez and Arellano (1998) [18].

The consistency result contrasts with those available for the structural equation setting, where 2SLS is inconsistent when the ratio of number of instruments to sample size tends to a positive constant [cf., Kunitomo (1980), Morimune (1983), Bekker (1994)]. Here the number of instruments, which is given by $T(T-1)/2$, increases very fast and yet consistency is obtained. The intuition for this result is that in our context as

[18] Here, for notational convenience, we assume that y_{i0} is also observed, so that the effective number of time series observations will be $T+1$.

T tends to infinity the "simultaneity bias" tends to zero, and so closeness of GMM1 or LIML1 to OLS in orthogonal deviations (ie. within-groups) becomes a desirable property.

Note that when $T/N \to 0$ the fixed T results for GMM1 and LIML1 remain valid, but within-groups, although consistent, has an asymptotic bias in its asymptotic distribution (which would only disappear if $N/T \to 0$). However, when T/N tends to a positive constant, within-groups, GMM1 and LIML1 exhibit negative biases in their asymptotic distributions. The condition that $c > 2$ is not restrictive since GMM1 and LIML1 are only well defined for $(T-1)/N \leqslant 1$. Thus, for $T < N$ the GMM1 bias is always smaller than the within-groups bias, and the LIML1 bias is smaller than the other two.

Another interesting feature is that the three estimators are asymptotically efficient in the sense of attaining the same asymptotic variance as the within-groups estimator as $T \to \infty$. However, Alvarez and Arellano (1998) show that the standard formulae for fixed T estimated variances of GMM1 and LIML1, which depend on the variance of the fixed effect, remain consistent estimates of the asymptotic variances as $T \to \infty$.

These results provide some theoretical support for LIML1 over GMM1. They also illustrate the usefulness of understanding the properties of panel data estimators as the time series information accumulates, even for moderate values of T: in a fixed T framework, GMM1 and LIML1 are asymptotically equivalent, but as T increases LIML1 has a smaller asymptotic bias than GMM1.

The crude GMM estimator in first differences. Alvarez and Arellano (1998) also show that the crude GMM estimator (CIV) that neglects the autocorrelation in the first differenced errors (ie., one-step GMM in first-differences with weight matrix equal to $(Z'Z)^{-1}$) is inconsistent as $T/N \to c > 0$, despite being consistent for fixed T. The result is:

$$\widehat{\alpha}_{\text{CIV}} \xrightarrow{p} \alpha - \frac{(1+\alpha)}{2}\left(\frac{c}{2-(1+\alpha)(2-c)/2}\right). \tag{93}$$

The intuition for this result is that the "simultaneity bias" of OLS in first differences (unlike the one for orthogonal deviations) does not tend to zero as $T \to \infty$. Thus, for fixed T the IV estimators in orthogonal deviations and first differences are both consistent, whereas as T increases the former remains consistent but the latter is inconsistent. Moreover, notice that the bias may be qualitatively relevant. Standard fixed-T large-N GMM theory would just describe the CIV estimator as being asymptotically less efficient than GMM1 as a consequence of using a non-optimal choice of weighting matrix.

4. Nonlinear panel data models

The ability to difference out the individual specific effect as was done in the previous sections relies heavily on the linear or multiplicative way in which it entered the model.

Many simple cross sectional models have a constant that does not enter in this way. This is for example true for all the limited dependent variable models discussed in Chapters 9 and 10 of Amemiya (1985). Introducing an individual specific effect as an individual specific constant in those models therefore results in models that cannot be estimated by the methods discussed so far. As will be seen in the following sections, the currently available methods for dealing with these models, rely on insights that are model-specific and that do not always seem to be useful for similar, but slightly different models. The main exception to this is the conditional maximum likelihood approach which has been used to construct estimators for some exponential family models. We discuss this method in the next section.

Unfortunately, there are many models for which it is not possible to use the conditional likelihood approach to eliminate the individual specific effect. For some of those models, alternative appoaches have been developed. In Sections 6 and 7, we will review some of the progress that has been made in the area of estimation of limited dependent variable models with individual-specific, “fixed”, effects[19]. This literature is closely related to the literature that deals with estimation of semiparametric limited dependent variables models, in that it is usually not necessary to specify a parametric form for the distribution of the underlying errors. The models are also semiparametric in the sense that the distribution of the individual specific effects conditional on the explanatory variable, is left unspecified. It is therefore not surprising that there is a close relationship between some of the approaches that are discussed here, and some approaches that have been taken to estimation of semiparametric limited dependent variables models. Indeed, in some cases the estimators for the panel data models have preceded the “corresponding” estimators for the cross sectional models.

The main limitation of much of the literature on nonlinear panel data methods, is that it is assumed that the explanatory variables are strictly exogenous in the sense that some assumptions will be made on the errors conditional on all (including future) values of the explanatory variables. As was pointed out earlier in this chapter, many of the recent advances in estimation of linear panel data models have focused on relaxing this assumption. In Section 8, we will discuss how some of the methods can be generalized to allow for lagged dependent variables, but at this point very little is known about estimation of nonlinear panel data models with predetermined explanatory variables.

The discussion of nonlinear panel data models in the next three sections will focus entirely on standard nonlinear econometric models in which the parameter that is usually interpreted as an intercept, is allowed to be individual specific. This seems like a natural first step in understanding the value and limitations of panel data when the model of interest is nonlinear. However, it is clear that knowing the “parameters

[19] Even though one often imagines a random sample of individuals, and hence random draws of the individual specific effects, it is customary to call the effect “fixed” when no assumptions are made on its relationship with other explanatory variables. A random effect is one which has been modelled in some manner.

of interest" in the models discussed below does not always allow one to infer all the quantities of interest. For example, in the fixed effects logit model below, knowing β will not allow one to infer the effect of one of the explanatory variables on the probability distribution of the dependent variable, although knowing the vector of β's will allow one to infer the relative effects of the explanatory variables. This problem is due to the semiparametric nature of the nonlinear models considered here, and is not particular to panel data. On the other hand, if the censoring in Equation (103) below is due to top – or bottom – coding of the true dependent variable of interest, then the interpretation of the parameters of the censored regression model is exactly the same as the interpretation of the parameters of a linear panel data model. The same can sometimes be said for the selection models discussed below.

Another limitation of most of the discussion here is that it focuses on the extreme case where no assumptions are made on the relationship between the individual specific effect and the explanatory variables. Whether a more "random" effects approach where some assumptions are made on how the distribution of this effect depends on the explanatory variables is more useful, depends on the context (and one's taste). In section 9 we briefly discuss some recent advances in this area. We devote much less space to that topic because many of the new developments there are by-products of developments in other areas of econometrics. For example, recent developments in Bayesian econometrics and in simulation-based inference have implications for nonlinear random effects panel data models, but the main new insights are more general, and not really tied to panel data.

5. Conditional maximum likelihood estimation

In a static linear model, one can justify treating the individual specific effects as parameters to be estimated by reference to the Frisch–Waugh Theorem: OLS (or normal maximum likelihood) on individual specific dummy variables is numerically equivalent to OLS on deviations from means. This means that including individual specific dummies yields a consistent estimator of the slope parameters (as n goes to infinity), even though the number of parameters is also going to infinity. Unfortunately, as was pointed in the classic paper by Neyman and Scott (1948), it is generally not the case that the maximum likelihood estimator will retain its nice asymptotic properties when the number of parameters is allowed to increase with sample size. This is for example seen by considering the maximum likelihood estimator of the variance in a static linear panel data model with normal errors: because the maximum likelihood estimator does not make the degrees-of-freedom correction, it will be inconsistent if the number of parameters is of order n.

Conditional maximum likelihood estimation is a method which, when it is applicable, can be used to construct consistent estimators of panel data models in the presence of individual specific effects. The idea is as follows. Suppose that a random variable, y_{it}, has distribution $f(\cdot; \theta, \alpha_i)$ where θ is the parameter of interest and is common for all i, whereas α_i is a nuisance parameter which is allowed to differ across i. A sufficient statistic, T_i, for α_i is a function of the data such that the distribution of the data given T_i does not depend on α_i. However, it might well depend on θ. If that is the case, then one can estimate θ by maximum likelihood using the conditional distribution of the data given the sufficient statistics. Andersen (1970) proved that the resulting estimator is consistent and asymptotically normal under appropriate regularity conditions. In the two subsections below, we give examples of how the conditional maximum likelihood estimator can be used to construct estimators of the panel data logit and the panel data Poisson regression models.

The problem with conditional maximum likelihood estimation as a general prescription for constructing estimators of nonlinear panel data models is that it is not always possible to find sufficient statistics such that the conditional distribution of the data conditional on the sufficient statistic will depend on θ. This is the case for many of the nonlinear models used in econometrics.

5.1. Conditional maximum likelihood estimation of logit models

The simplest interesting nonlinear model for which the conditional likelihood approach works, is the "textbook" logit model studied in Rasch (1960, 1961). With two time periods and an individual specific constant we have,

$$y_{it} = 1\left\{x_{it}\beta + \alpha_i + \varepsilon_{it} \geqslant 0\right\} \qquad t = 1,2,\ i = 1, \dots, n$$

where ε_{i1} and ε_{i2} are independent and logistically distributed, conditional on α_i, x_{i1}, x_{i2}. It follows that

$$\Pr\left(y_{it} = 1 | x_{i1}, x_{i2}, \alpha_i\right) = \frac{\exp\left(x_{it}\beta + \alpha_i\right)}{1 + \exp\left(x_{it}\beta + \alpha_i\right)}. \tag{94}$$

In this case it is easy to see how the conditional likelihood approach "eliminates" the individual specific effect. Define events A and B by $A = \{y_{i1} = 0, y_{i2} = 1\}$ and $B = \{y_{i1} = 1, y_{i2} = 0\}$. It is then an easy exercise to show that

$$\begin{aligned}\Pr\left(y_{i1} = 0, y_{i2} = 1 |\, y_{i1} + y_{i2} = 1, x_{i1}, x_{i2}, \alpha_i\right) &= \Pr\left(A | A \cup B, x_{i1}, x_{i2}, \alpha_i\right) \\ &= \frac{1}{1 + \exp\left(\left(x_{i1} - x_{i2}\right)\beta\right)}.\end{aligned} \tag{95}$$

In words, if we restrict the sample to the observations for which y_{it} changes, then the individual specific effects do not enter the distribution of (y_{i1}, y_{i2}) given $(x_{i1}, x_{i2}, \alpha_i)$ and the distribution of y_{i1} given (x_{i1}, x_{i2}) has the form of a logit model with explanatory variable $x_{i1} - x_{i2}$ and coefficient β. Intuitively, the implication is that if we restrict the sample to the observations for which y_{it} changes over time, then β can be estimated by estimating a logit in the restricted sample without having to specify the distribution of the individual specific effects. In a sense, conditioning on $y_{i1} + y_{i2} = 1$ has the same effect as differencing the data in a linear panel data model.

More generally, if there are $T > 2$ observations for each individual, the conditional distribution of $(y_{i1}, \dots, y_{it})$ given $\sum_{t=1}^{T}, y_{it}$ is

$$P\left(y_{i1}, \dots, y_{it} \,\Big|\, \sum_{t=1}^{T} y_{it}, x_{i1}, \dots, x_{it}, \alpha_i\right) = \frac{\exp\left(\sum_{t=1}^{T} y_{it} x_{it} \beta\right)}{\sum_{(d_1, \dots, d_t) \in B} \exp\left(\sum_{t=1}^{T} d_t x_{it} \beta\right)}, \tag{96}$$

where B is the set of all sequences of zeros and ones that have $\sum_{t=1}^{T} d_{it} = \sum_{t=1}^{T} y_{it}$. Formally this means that $\sum_{t=1}^{T} y_{it}$ is a sufficient statistic for α_i, and the implication is that one can use Equation (96) to estimate β. Chamberlain (1980) generalized Equation (96) by deriving the conditional likelihood for the multinomial logit model.

When T is large, the number of terms in the denominator of Equation (96) will be large, and and it can be computationally burdensome to calculate the conditional maximum likelihood estimator. In that case one can estimate β by applying the logic leading to Equation (95) to all pairs of observations for a given individual. In other words, one can maximize

$$\sum_{i=1}^{n}\left(\sum_{s<t} \log\left(\frac{\exp\left(y_{it}\left(x_{it} - x_{is}\right)\beta\right)}{1 + \exp\left(\left(x_{it} - x_{is}\right)\beta\right)}\right)\right).$$

Unless $T = 2$, this objective function is not a (log-)likelihood, and it will generally be less efficient than the conditional maximum likelihood estimator. The asymptotic distribution of the estimator can be found by noting that it is an extremum estimator.

5.2. Poisson regression models

The Poisson regression model with individual specific constants provides another example in which the conditional maximum likelihood estimator can be used. This is a special case of the multiplicative model discussed earlier. For simplicity, consider the case where there are two observations for each individual:

$$y_{it} \sim \text{po}\left(\exp(\alpha_i + x_{it}\beta)\right) \qquad t = 1, 2 \qquad i = 1, \dots, n. \tag{97}$$

One way to understand why the conditional likelihood approach will work in this model, is to recall that if two independent random variables are both Poisson distributed with means μ_1 and μ_2, respectively, then the distribution of one of them given the sum, has a binomial distribution with probability parameter $\frac{\mu_1}{\mu_1 + \mu_2}$ and trial parameter given by the sum of the two random variables. It therefore follows that if y_{i1} and y_{i2} are drawn from Equation (97) and we restrict attention to the observations for which $y_{i1} + y_{i2} = K$ (say), then $y_{i1} \sim \text{bi}\left(K, \frac{\exp(x_{i1}\beta)}{\exp(x_{i1}\beta) + \exp(x_{i2}\beta)}\right)$. Since this distribution does not involve the individual specific effects, it can be used to make inference about β. For example, one could estimate β by maximizing

$$L = \sum_i -y_{i1} \ln\left(1 + \exp((x_{i2} - x_{i1})b)\right) - y_{i2} \ln\left(1 + \exp((x_{i1} - x_{i2})b)\right)$$

[see, for example, Hausman, Hall and Griliches (1984)].

Recent papers by Blundell, Griffith and Windmeijer (1997), and Lancaster (1997) have pointed out that for the Poisson regression model (97), the conditional maximum likelihood estimator is identical to the maximum likelihood estimator of β based on maximizing the likelihood function for Equation (97) over b and all the individual specific effects, α_i.

6. Discrete choice models with "fixed" effects

Manski (1987) made the first successful attempt of consistently estimating a nonlinear panel data model with individual specific "fixed" effects in a situation in which the conditional maximum likelihood approach cannot be applied. His estimator is based on the maximum score estimator [see Manski (1975)] for the binary choice model

$$y_i = 1\left\{x_i\beta + \varepsilon_i \geqslant 0\right\}. \tag{98}$$

Since $P(y_i = 1 | x_i) = F_{-\varepsilon_i|x_i}(x_i\beta)$ it follows that if $\text{Median}(\varepsilon_i|x_i) = 0$ (uniquely), then observations with $x_i\beta > 0$ will have probabilities greater than $\frac{1}{2}$ and observations with $x_i\beta < 0$ will have probabilities less than $\frac{1}{2}$. In other words,

$$\operatorname{sgn}\left(\Pr(y_i = 1|x_i) - \Pr(y_i = 0|x_i)\right) = \operatorname{sgn}(x_i\beta).$$

Under mild regularity conditions, this implies that $E[\operatorname{sgn}(2y_i - 1)\operatorname{sgn}(x_ib)]$ is uniquely maximized at $b = \beta$, and the analogy principle therefore suggests estimating β by

$$\widehat{\beta} = \arg\max_b \sum_{i=1}^{n} \operatorname{sgn}(2y_i - 1)\operatorname{sgn}(x_ib).$$

Under mild conditions, this estimator is consistent [see Manski (1985)], but it does not converge at rate $\sqrt{n}$ and it is not asymptotically normal [see Cavanagh (1987) and Kim and Pollard (1990)][20].

[20] Under assumptions that are slightly stronger than Manski's, Horowitz (1992) proposed a smoothed version of the maximum score estimator which does have an asymptotic normal distribution, although the rate is, again, slower than $\sqrt{n}$. The rate of convergence of Horowitz's estimator depends on the assumed degree of smoothness of the distribution of the explanatory variables.

The insight behind Manski's (1987) estimator of the ("non-logit") binary choice model with individual specific effects, is that under mild conditions, exactly the same conditioning that leads from the logit model with individual specific fixed effects (94) to a logit model without the individual specific "fixed" effects, (95), will also lead from the model

$$y_{it} = 1\left\{x_{it}\beta + \alpha_i + \varepsilon_{it} \geqslant 0\right\} \quad t = 1, 2;\ i = 1, \ldots, n, \tag{99}$$

to a model in which the maximum score estimator can be applied. The key assumption is that the distribution of ε_{it} is stationary, in the sense that ε_{i1} and ε_{i2} are identically distributed conditional on $(x_{i1}, x_{i2}, \alpha_i)$. With this assumption, Manski showed that

$$\Pr\left(y_{i2} = 1 | x_{i1}, x_{i2}, y_{i1} + y_{i2} = 1\right) \lesseqgtr 1/2,$$

depending on whether

$$\left(x_{i2} - x_{i1}\right)\beta \lesseqgtr 0.$$

The intuition for this result is simple. If the distribution of $-\varepsilon_{i1}$ (and $-\varepsilon_{i2}$) for individual i is $F_i(\cdot)$, then the probability that $y_{it} = 1$ for individual i is $F_i(x_{it}\beta + \alpha_i)$; this means that for a given individual, higher values of $x_{it}\beta$ are more likely to be associated with $y_{it} = 1$.

Mimicking Manski (1975), this suggests a conditional maximum score estimator defined by

$$\widehat{\beta} = \arg\max_b \sum_{i=1}^{n} \operatorname{sgn}\left(y_{i2} - y_{i1}\right) \operatorname{sgn}\left(\left(x_{i2} - x_{i1}\right)b\right). \tag{100}$$

If the panel is of length longer than 2, one can estimate β by considering all pairs of observations

$$\widehat{\beta} = \arg\max_b \sum_{i=1}^{n} \sum_{s<t} \left(\operatorname{sgn}\left(y_{is} - y_{it}\right) \operatorname{sgn}\left(\left(x_{is} - x_{it}\right)b\right)\right). \tag{101}$$

As was the case for the cross sectional maximum score estimator, this estimator will be consistent under mild regularity conditions. In particular, compared to the logit model considered earlier, it not only leaves the distribution of the errors unspecified, but it also allows for general serial correlation and heteroskedasticity across individuals (but not over time). However, the estimator is not $\sqrt{n}$ consistent, and not asymptotically normal[21].

[21] Kyriazidou (1995) and Charlier, Melenberg and van Soest (1995) have shown that the same trick used by Horowitz (1992) to modify the maximum score estimator can be used to modify the conditional maximum score estimator. This results in a smoothed conditional maximum score estimator which does have an asymptotic normal distribution, although the rate is, again, slower than $\sqrt{n}$.

Since on one hand, Manski's estimator is not $\sqrt{n}$ consistent, but makes very weak assumptions on the errors, and on the other hand assuming a logistic distribution on the errors leads to a $\sqrt{n}$ consistent and asymptotically normal estimator, it is natural to ask whether there are alternative assumptions on the errors that lead to a situation where it is possible to estimate the β-vector at the usual $\sqrt{n}$ rate. Perhaps surprisingly, the answer to that question seems to be negative. Subject to weak regularity conditions Chamberlain (1993), showed that even if ε_{it} in Equation (99) are i.i.d. with known distribution and independent of $(x_{i1}, x_{i2}, \alpha_i)$, β can be estimated $\sqrt{n}$ consistently only in the logit case.

It is clear that scale normalizations are needed in each period in order for β in Equation (99) to be identified. Both the logit version of Equation (99) and Manski's treatment impose such scale normalizations. In the logit case, this normalization comes from the variance of the logistic distribution. In Manski's case it is through a scale normalization on β and through the assumption that the errors are identically distributed in the two time periods. In addition to these scale normalizations, the estimators of Equation (99) also assume that the effect of the fixed effect is the same in the two periods. This is in contrast to the linear model in which it is possible to estimate time specific coefficients (factor loadings) on the fixed effect. It is clear that the logic behind the two estimators of the binary choice panel data model discussed here would break down with such factor loadings, but it is less clear whether they would make the model unidentified.

7. Tobit-type models with "fixed" effects

7.1. Censored regression models

The censored regression model is given by

$$\begin{aligned} y_i^* &= x_i\beta + \varepsilon_i \\ y_i &= \max\{y_i^*, c\} \end{aligned} \tag{102}$$

In text-book treatments, c is usually 0. Note that for $c = -\infty$, Equation (102) becomes the linear regression model, and that one can change the max to a min by a simple change of sign. The censored regression model has been used in many different contexts. In some, c is the lowest possible value that some economic variable can take, and y^* is the desired level of that variable in the absence of this constraint. In other cases, the censoring is induced by the way the data is constructed. For example, earnings variables are sometimes top-coded for confidentiality reasons.

In a panel data context, the censored regression model may be described by

$$\begin{aligned} y_{it}^* &= x_{it}\beta + \alpha_i + \varepsilon_{it} \\ y_{it} &= \max\{y_{it}^*, c\} \end{aligned} \tag{103}$$

This model was introduced by Heckman and MaCurdy (1980) in the context of female labor supply.

Because the individual specific effect α_i does not enter linearly or multiplicatively, it is not possible to "difference" it out as was the case for the linear regression model, and it is also unclear under what conditions a conditional likelihood approach can be used to eliminate α_i. Honoré (1992) proposed a different approach to estimating β in this model. The motivation for the estimators given below is different from that in Honoré (1992) because we want to motivate a larger class of estimators. Honoré (1992) also considered estimation of the truncated version of the model. The latter is less interesting and will not be discussed here.

The idea behind the estimator in Honoré (1992) is to artificially censor the dependent variable in such a way that the individual specific effect can be differenced away. This is similar to the approach in Powell (1986) who artificially censored the dependent variable in a cross sectional censored regression model, in such a way that the moment conditions for OLS apply. Specifically, one can define pairs of "residuals" that depend on the individual specific effect in exactly the same way. Intuitively, this implies that differencing the residuals will eliminate the fixed effects.

Define

$$\upsilon_{ist}(b) = \max\{y_{is}, c + (x_{is} - x_{it})\,b\} - \max\{c, c + (x_{is} - x_{it})\,b\}$$

At $b = \beta$, we have

$$\begin{aligned}\upsilon_{ist}(\beta) &= \max\{y_{is}, c + (x_{is} - x_{it})\,\beta\} - \max\{c, c + (x_{is} - x_{it})\,\beta\}\\ &= \max\{\alpha_i + \varepsilon_{is}, c - x_{is}\beta, c - x_{it}\beta\} - \max\{c - x_{is}\beta, c - x_{it}\beta\}\end{aligned}$$

The key observation is that $\upsilon_{ist}(\beta)$ is symmetric in s and t. Therefore, if ε_{it}, $t = 1, \ldots, T$, are independent and identically distributed conditional on (x_i, α_i), where x_i denotes all the explanatory variables for individual i, then $\upsilon_{ist}(\beta)$ and $\upsilon_{its}(\beta)$ are independent and identically distributed (conditional on (x_i, α_i)). This means that any function of $\upsilon_{ist}(\beta)$ minus the same function of $\upsilon_{its}(\beta)$ will be symmetrically distributed around 0. We therefore have the conditional moment condition

$$E\left[(\xi(\psi(\upsilon_{its}(\beta)) - \psi(\upsilon_{ist}(\beta))))\middle| x_i, \alpha_i\right] = 0, \tag{104}$$

for any increasing function $\psi(\cdot)$ and any increasing and odd function $\xi(\cdot)$, provided that the expectations are well-defined. The reason why $\psi(\cdot)$ and $\xi(\cdot)$ are assumed to be increasing will become clear shortly.

One could in principle consider estimation of β on the basis of Equation (104). One problem with this is that although β satisfies Equation (104), it does not follow from the previous discussion that there are no other values of the parameter that also satisfy Equation (104). However, Equation (104) implies

$$E[(\xi(\psi(\upsilon_{its}(\beta)) - \psi(\upsilon_{ist}(\beta))))(x_{it} - x_{is})] = 0, \tag{105}$$

which has the form

$$E\left[r\left(y_{is}, y_{it}, (x_{is}-x_{it})\beta\right)(x_{is}-x_{it})\right]=0, \tag{106}$$

where $r(\cdot,\cdot,\cdot)$ is a monotone function of its third argument, because of the assumption that $\psi(\cdot)$ and $\xi(\cdot)$ are increasing[22]. By integrating $r(\cdot)$ with respect to its third argument, one can typically turn Equation (106) into the first order condition for a convex minimization problem of the form

$$\min_b E\left[R\left(y_{is}, y_{it}, (x_{is}-x_{it})b\right)\right]. \tag{107}$$

The parameter β can then be estimated by minimizing a sample analog of Equation (107). It follows from standard results about extremum estimators that the resulting estimator will be consistent and $\sqrt{n}$ asymptotically normal.

For example, with $\xi(d)=\psi(d)=d$, $c=0$ and $T=2$, the function to be minimized in (107) becomes

$$\begin{aligned} E\Big[&(\max\{y_{i1},\Delta x_i b\}-\max\{y_{i2},-\Delta x_i b\}-\Delta x_i b)^2 \\ &+2\cdot 1\{y_{i1}<\Delta x_i b\}(\Delta x_i b-y_{i1})y_{i2}+2\cdot 1\{y_{i2}<-\Delta x_i b\}(-\Delta x_i b-y_{i2})y_{i1}\Big], \end{aligned}$$

which suggests estimating β by

$$\begin{aligned} \widehat{\beta}=\arg\min_b \sum_{i=1}^{n} &(\max\{y_{i1},\Delta x_i b\}-\max\{y_{i2},-\Delta x_i b\}-\Delta x_i b)^2 \\ &+2\cdot 1\{y_{i1}<\Delta x_i b\}(\Delta x_i b-y_{i1})y_{i2} \\ &+2\cdot 1\{y_{i2}<-\Delta x_i b\}(-\Delta x_i b-y_{i2})y_{i1}. \end{aligned}$$

Letting $\xi(d)=\operatorname{sign}(d)$ and $\psi(d)=d$, results in the estimator

$$\begin{aligned} \widehat{\beta}\ &=\arg\min_b \sum_{i=1}^{n}(1-1\{y_{i1}\leqslant \Delta x_i b, y_{i2}\leqslant 0\}) \\ &\cdot(1-1\{y_{i2}\leqslant -\Delta x_i b, y_{i1}\leqslant 0\})\left|y_{i1}-y_{i2}-\Delta x_i b\right|. \end{aligned}$$

These are the estimators discussed in detail in Honoré (1992). Honoré and Kyriazidou (2000b) discuss estimators defined by a general $\psi(d)$ and $\xi(d)=d$ as well as $\psi(d)=d$ and general $\xi(d)$. The case with panels of length $T>2$ can be dealt with by considering all pairs of time periods s and t, as in Equation (101).

[22] $\upsilon_{ist}(b)=\max\{y_{is}, c+(x_{is}-x_{it})b\}-\max\{c, c+(x_{is}-x_{it})b\}$ is monotone in $(x_{is}-x_{it})b$ because $y_{is}\geqslant c$. It therefore follows that $\xi(\psi(\upsilon_{its}(b))-\psi(\upsilon_{ist}(b)))$ depends on b only through $(x_{is}-x_{it})b$ and that it is monotone in $(x_{is}-x_{it})b$.

The moment condition (105) was derived from the assumption that ε_{is} and ε_{it} are independent and identically distributed conditional on (x_i, α_i). This assumption is stronger than necessary. To see why, assume the conditional exchangeability assumption that $(\varepsilon_{is}, \varepsilon_{it})$ is distributed like $(\varepsilon_{it}, \varepsilon_{is})$ conditional on (x_i, α_i). This implies that $(\psi(\upsilon_{ist}(\beta)), \psi(\upsilon_{its}(\beta)))$ is distributed like $(\psi(\upsilon_{its}(\beta)), \psi(\upsilon_{ist}(\beta)))$, which in turn implies that $\psi(\upsilon_{ist}(\beta)) - \psi(\upsilon_{its}(\beta))$ is symmetrically distributed around 0 (all conditional on (x_i, α_i)). The moment condition (105) then follows.

The exchangeability condition is useful because it yields symmetry of $\psi(\upsilon_{ist}(\beta)) - \psi(\upsilon_{its}(\beta))$, which then yields the moment condition for *any* choice of the odd function ξ. On the other hand, if ξ is the identity function, then the moment condition follows if $\psi(\upsilon_{ist}(\beta))$ is distributed like $\psi(\upsilon_{its}(\beta))$, which is implied by ε_{is} and ε_{it} being identically distributed. In other words, the stationarity assumption that was the key to Manski's estimator for the panel data binary choice model, is also the key to the class of estimators for the panel data censored regression model based on the moment condition (106) (and the minimization problem (107)) with $\xi(d) = d$, whereas the larger class of estimators based on Equation (106) with general ξ seems to require the stronger assumption that ε_{is} and ε_{it} are exchangeable.

7.2. Type 2 Tobit model (sample selection model)

Kyriazidou (1997) studied the more complicated model

$$\begin{aligned} y_{1it}^* &= x_{1it}\beta_1 + \alpha_{1i} + \varepsilon_{1it}, \\ y_{2it}^* &= x_{2it}\beta_2 + \alpha_{2i} + \varepsilon_{2it}, \end{aligned}$$

where we observe:

$$y_{1it} = 1\{y_{1it}^* > 0\} \tag{108}$$

$$y_{2it} = \begin{cases} y_{2it}^* & \text{if } y_{1it} = 1 \\ 0 & \text{otherwise} \end{cases}. \tag{109}$$

This is a panel data version of the sample selection model that Amemiya (1985) calls the Type 2 Tobit Model.

It is clear that β_1 can be estimated by one of the methods for estimation of discrete choice models with individual specific effects discussed earlier. Kyriazidou's insight into estimation of β_2 combines insights from the literature on estimation of semiparametric sample selection models with the idea of eliminating the individual specific effects by first-differencing the data. Specifically, to difference out the individual specific effects α_{2i}, one must restrict attention to observations for which

y_{2it}^{*} is observed. With this "sample selection", the mean of the error term in period t is

$$\lambda_{it} = E\left(\varepsilon_{2it} \mid \varepsilon_{1it} > -x_{1it}\beta_1 - \alpha_{1i}, \varepsilon_{1is} > -x_{1is}\beta_1 - \alpha_{1i}, \zeta_i\right),$$

where $\zeta_i = (x_{1is}, x_{2is}, x_{1it}, x_{2it}, \alpha_{i1}, \alpha_{i2})$. The key observation in Kyriazidou (1997) is that if $(\varepsilon_{1it}, \varepsilon_{2it}, \varepsilon_{1is}, \varepsilon_{2is})$ and $(\varepsilon_{1is}, \varepsilon_{2is}, \varepsilon_{1it}, \varepsilon_{2it})$ are identically distributed (conditional on $(x_{1is}, x_{2is}, x_{1it}, x_{2it}, \alpha_{i1}, \alpha_{i2})$), then for an individual i, who has $x_{1it}\beta_1 = x_{1is}\beta_1$,

$$\begin{aligned} \lambda_{it} &= E\left(\varepsilon_{2it} \mid \varepsilon_{1it} > -x_{1it}\beta_1 - \alpha_{1i}, \varepsilon_{1is} > -x_{1is}\beta_1 - \alpha_{1i}, \zeta_i\right) \\ &= E\left(\varepsilon_{2is} \mid \varepsilon_{1is} > -x_{1is}\beta_1 - \alpha_{1i}, \varepsilon_{1it} > -x_{1it}\beta_1 - \alpha_{1i}, \zeta_i\right) \\ &= \lambda_{is}. \end{aligned} \tag{110}$$

This implies that for individuals with $x_{1it}\beta_1 = x_{1is}\beta_1$, the same first differencing that will eliminate the fixed effect will also eliminate the effect of sample selection. This suggests a two-step estimation procedure similar to Heckman's (1976, 1979) two-step estimator of sample selection models: first estimate β_1 by one of the methods discussed earlier, and then, secondly, estimate β_2 by applying OLS to the first differences, but giving more weight to observations for which $(x_{1it} - x_{1is})\widehat{\beta}_1$ is close to zero:

$$\begin{aligned} \widehat{\beta}_2 = &\left[\sum_{i=1}^{n}\sum_{s<t}(x_{2it} - x_{2is})'(x_{2it} - x_{2is})K\left(\frac{(x_{1it} - x_{1is})\widehat{\beta}_1}{h_n}\right)y_{1it}y_{1is}\right]^{-1} \\ &\times\left[\sum_{i=1}^{n}\sum_{s<t}(x_{2it} - x_{2is})'(y_{2it} - y_{2is})K\left(\frac{(x_{1it} - x_{1is})\widehat{\beta}_1}{h_n}\right)y_{1it}y_{1is}\right] \end{aligned}$$

where K is a kernel and h_n is a bandwidth which shrinks to zero as the sample size increases. Kyriazidou showed that the resulting estimator is $\sqrt{nh_n}$-consistent and asymptotically normal.

Kyriazidou's estimator is closely related to the estimator proposed by Powell (1987). That paper considered a cross sectional sample selection model and applied the argument leading to Equation (110) to all pairs of observations i and j.

7.3. Other Tobit-type models

As pointed out in Honoré and Kyriazidou (2000b), the estimators proposed in Honoré (1992) and Kyriazidou (1997) can be modified fairly trivially to cover the other

Tobit-type models discussed in Amemiya (1985). Consider for example, the Type 3 Tobit model with individual-specific effects,

$$y_{1it}^* = x_{1it}\beta_1 + \alpha_{1i} + \varepsilon_{1it}$$

$$y_{2it}^* = x_{2it}\beta_2 + \alpha_{2i} + \varepsilon_{2it}$$

$$y_{1it} = \begin{cases} y_{1it}^* & \text{if } y_{1it}^* > 0 \\ 0 & \text{if } y_{1it}^* \leqslant 0 \end{cases}$$

$$y_{2it} = \begin{cases} y_{2it}^* & \text{if } y_{1it}^* > 0 \\ 0 & \text{if } y_{1it}^* \leqslant 0 \end{cases}.$$

In that model, the event

$$E = \{y_{1is} > \max\{0, (x_{1is} - x_{1it})\beta_1\},\ y_{1it} > \max\{0, (x_{1it} - x_{1is})\beta_1\}\}$$

is the same as the event

$$\begin{aligned}&\big\{\varepsilon_{1is} > \max\{-x_{1is}\beta_1 - \alpha_{1i}, -x_{1it}\beta_1 - \alpha_{1i}\},\\ &\varepsilon_{1it} > \max\{-x_{1is}\beta_1 - \alpha_{1i}, -x_{1it}\beta_1 - \alpha_{1i}\}\big\}.\end{aligned}$$

With the exchangeability assumption that $(\varepsilon_{1it}, \varepsilon_{2it}, \varepsilon_{1is}, \varepsilon_{2is})$ and $(\varepsilon_{1is}, \varepsilon_{2is}, \varepsilon_{1it}, \varepsilon_{2it})$ are identically distributed (conditional on $(x_{1is}, x_{2is}, x_{1it}, x_{2it}, \alpha_{i1}, \alpha_{i2})$)

$$\varepsilon_{1is} - \varepsilon_{1it} = (y_{2is} - y_{2it}) - (x_{2is} - x_{2it})\beta_2,$$

is symmetrically distributed around 0 conditional on E and conditional on $(x_{1is}$, x_{2is}, x_{1it}, x_{2it}, α_{i1}, $\alpha_{i2})$. This suggests a two-step approach, where the first step is estimation of β_1 by one of the estimators of the panel data censored regression, and the second step estimates β_2 by

$$\begin{aligned}\widehat{\beta}_2 = \arg\min_b \sum_i \sum_{s<t} 1 \Big\{y_{1is} > \max\{0, (x_{1is} - x_{1it})\widehat{\beta}_1\},\ y_{1it} > \max\{0, (x_{1it} - x_{1is})\widehat{\beta}_1\}\Big\}\\ \cdot\, \Xi\left((y_{is} - y_{it}) - (x_{is} - x_{it})\, b\right),\end{aligned}$$

where Ξ is some symmetric loss function such as $\Xi(d) = d^2$ or $\Xi(d) = |d|$.

The Type 3 Tobit model was also considered by Ai and Chen (1992) who presented moment conditions similar to those implied by the two-step estimator above, although they derived their conditions under the assumption that the errors are independent over time.

It is also straightforward to consider panel data versions of Amemiya's Type 4 and Type 5 Tobit Models. Let

$$\begin{aligned} y_{1it}^* &= x_{1it}\beta_1 + \alpha_{1i} + \varepsilon_{1it}, \\ y_{2it}^* &= x_{2it}\beta_2 + \alpha_{2i} + \varepsilon_{2it}, \\ y_{3it}^* &= x_{3it}\beta_3 + \alpha_{3i} + \varepsilon_{3it}. \end{aligned}$$

In the Type 4 Tobit model we observe $(y_{1it}, y_{2it}, y_{3it})$ from:

$$y_{1it} = \max\left\{0, y_{1it}^*\right\}, \tag{111}$$

$$y_{2it} = \begin{cases} y_{2it}^* & \text{if } y_{1it}^* > 0 \\ 0 & \text{otherwise} \end{cases}, \tag{112}$$

$$y_{3it} = \begin{cases} y_{3it}^* & \text{if } y_{1it}^* \leqslant 0 \\ 0 & \text{otherwise} \end{cases}, \tag{113}$$

and we can estimate the parameters of this model by considering Equations (111) and (112) as a Type 3 Tobit model and Equations (111) and (113) as a Type 2 sample selection model.

In the Type 5 Tobit model we observe $(y_{1it}, y_{2it}, y_{3it})$ from:

$$y_{1it} = 1\left\{y_{1it}^* > 0\right\}, \tag{114}$$

$$y_{2it} = \begin{cases} y_{2it}^* & \text{if } y_{1it} = 1 \\ 0 & \text{otherwise} \end{cases}, \tag{115}$$

$$y_{3it} = \begin{cases} y_{3it}^* & \text{if } y_{1it} = 0 \\ 0 & \text{otherwise} \end{cases}, \tag{116}$$

and we can treat the two outcome Equations (115) and (116) separately and apply Kyriazidou's (1997) estimator to β_2 and β_3.

7.4. Monotone transformation models

Estimation of β in the cross sectional linear transformation model,

$$h(y_i) = x_i\beta + \varepsilon_i, \tag{117}$$

has been the topic of a large number of recent papers in econometrics and statistics. In this model, β is often considered the primary parameter of interest with h and the distribution of ε left unspecified except that $h(\cdot)$ is assumed to be monotone and ε independent of x. In some cases, h is assumed to be strictly monotone, whereas other papers do not require this, in which case Equation (117) contains both the binary discrete choice and the censored regression model as special cases. When h is assumed to be strictly monotone, one might think of Equation (117) as a generalization of the

Box–Cox model. It is clear that β can only be estimated up to scale, unless a scale normalization is imposed on $h(\cdot)$ or ε. In the following, we will therefore only be concerned with estimation of β up to scale.

In a recent paper, Abrevaya (1999) proposed an estimator of β in a fixed effects version of Equation (117),

$$h_t(y_{it}) = x'_{it}\beta + \alpha_i + \varepsilon_{it}, \tag{118}$$

where $h_t(\cdot)$ is assumed strictly increasing. His estimator is similar in spirit to that of Han (1987) for the cross sectional transformation model. The key insight in Abrevaya's paper is to difference across individuals in a given time period, rather than across time periods for a given individual,

$$h_t(y_{it}) - h_t(y_{jt}) = (x_{it} - x_{jt})'\beta + (\alpha_i - \alpha_j) + (\varepsilon_{it} - \varepsilon_{jt}).$$

Because h_t is strictly increasing,

$$\begin{aligned}\Pr(y_{it} &> y_{jt} \mid x_{it}, x_{is}, \alpha_i, x_{jt}, x_{js}, \alpha_j)\\ &= \Pr(\varepsilon_{jt} - \varepsilon_{it} < (x_{it} - x_{jt})'\beta + (\alpha_i - \alpha_j) \mid x_{it}, x_{is}, \alpha_i, x_{jt}, x_{js}, \alpha_j),\end{aligned}$$

where the motivation for conditioning of the explanatory variables in both time periods t and s, is that we will compare this probability in time period t to the same probability in time period s.

Assume that the errors are stationary (given the explanatory variables in all periods and given the fixed effect). This is the same assumption that was made for the discrete choice model and for the censored regression model. This assumption, combined with random sampling, implies that the distribution of $\varepsilon_{jt} - \varepsilon_{it}$ (given $(x_{it}, x_{is}, \alpha_i, x_{jt}, x_{js}, \alpha_j)$) is the same in the two periods. The right hand side of Equation (119) can then be written as $F_{ij}((x_{it} - x_{jt})'\beta + (\alpha_i - \alpha_j))$. On the other hand, by simple inspection it is clear that

$$\Delta x'_i\beta > \Delta x'_j\beta \Leftrightarrow (x_{it} - x_{jt})'\beta + (\alpha_i - \alpha_j) > (x_{is} - x_{js})'\beta + (\alpha_i - \alpha_j), \tag{119}$$

where $\Delta x = x_t - x_s$. Combining Equations (119) and (119) we then have[23]

$$\begin{aligned}\Delta x'_i\beta &> \Delta x'_j\beta \Rightarrow\\ \Pr(y_{it} &> y_{jt} \mid x_{is}, x_{it}, \alpha_i, x_{js}, x_{jt}, \alpha_j) > \Pr(y_{is} > y_{js} \mid x_{is}, x_{it}, \alpha_i, x_{js}, x_{jt}, \alpha_j).\end{aligned} \tag{120}$$

Equation (120) implies that the function

$$S(b) \equiv E\left[\operatorname{sign}\left(\left(\Delta x_i - \Delta x_j\right)' b\right)\left(1\left(y_{it} > y_{jt}\right) - 1\left(y_{is} > y_{js}\right)\right)\right], \tag{121}$$

[23] Some smoothness of the distribution of the errors is needed for the inequality between the probabilities to be strict.

is maximized at $b = \beta$. For the case where there are only two time periods, Abrevaya therefore proposed an estimator defined by maximizing the sample analog of Equation (121),

$$S_n(b) \equiv \binom{n}{2}^{-1} \sum_{i \neq j} \operatorname{sign}((\Delta x_i - \Delta x_j)' b)(1(y_{i2} > y_{j2}) - 1(y_{i1} > y_{j1})). \tag{122}$$

Abrevaya (1999) showed that his estimator is consistent and $\sqrt{n}$ asymptotically normal under appropriate regularity conditions. He also showed that although there are n^2 terms in the sum in Equation (122), it is possible to calculate the sum using $O(n \log(n))$ operations. The computational burden associated with the estimator is therefore much smaller that it appears. The case with $T > 2$ observations for each individual can again be dealt with by considering all pairs of time periods.

Abrevaya (2000) proposed an estimator for a model which is more general than Equation (118). That estimator is based on the same idea as Manski's (1985) maximum score estimator of the panel data binary choice estimator. As is the case for the maximum score estimator, it is possible to show that a smoothed version of Abrevaya's estimator is consistent and asymptotocally normal, although the rate of convergence is slower than $\sqrt{n}$.

7.5. *Nonparametric regression and fixed effects*

Porter (1997) introduced individual-specific additive effects in a nonparametric regression model by specifying

$$y_{it} = m_t(x_{it}) + \alpha_i + \varepsilon_{it}, \tag{123}$$

where ε_{it} has mean 0 conditional on all (past, current and future) values of the explanatory variables x_{it}. Porter noted that Equation (123) implies that the conditional mean of $y_{it} - y_{is}$ given (x_{it}, x_{is}) is $\ell(x_{it}, x_{is}) \equiv m_t(x_{it}) - m_s(x_{is})$. The latter can be estimated by standard techniques for nonparametric regression [see e.g., Härdle and Linton (1994)], and $m_t(\cdot)$ can then be recovered (except for an additive constant) by averaging ℓ over its second argument.

7.6. *Relationship with estimators for some cross sectional models*

The estimators for the panel data versions of the discrete choice model, the censored and truncated regression models, the sample selection model and the monotone transformation model all have "cousins" for the cross sectional versions of the models. The relationship is most easily understood by considering a simple cross sectional linear regression model where the observations consist of i.i.d. draws of

$$y_i = \alpha + x_i\beta + \varepsilon_i. \tag{124}$$

In this model, any two observations have the same intercept α. With some potential loss of information, one can therefore think of any two observations as if they are from a (static) linear panel data model with $T = 2$. This suggests forming all pairs of observations, and then estimating the slope-parameters β in Equation (124) by

$$\widehat{\beta} = \arg\min_b \sum_{i<j} \left(\left(y_i - y_j \right) - \left(x_i - x_j \right) b \right)^2 .$$

It is an easy exersice to show that this is nothing but the OLS estimator of β in the regression of Equation (124).

The same logic can be applied to nonlinear models. If the model under consideration is such that the parameter β can be estimated from a two-period panel by, say, some minimization problem

$$\widehat{\beta} = \arg\min_b \sum_i g\left(y_{i1}, y_{i2}, x_{i1}, x_{i2}, b \right),$$

then a cross sectional version of the model can be estimated by

$$\widehat{\beta} = \arg\min_b \sum_{i<j} g\left(y_i, y_j, x_i, x_j, b \right).$$

Honoré and Powell (1994) applied this insight to construct estimators for the cross sectional censored and truncated regression models based on the panel data estimators in Honoré (1992).

The panel data estimators for the discrete choice and sample selection models also have cross sectional versions. If Manski's (1987) estimator is applied to all pairs of observations from a cross sectional binary choice model, then the maximum rank correlation estimator of Han (1987) results (although his motivation was quite different and his estimator applies to a more general class of transformations models). Likewise, applying the logic behind Kyriazidou's (1997) estimator of the sample selection model to all pairs of observations in a cross sectional sample selection model results in the estimator proposed by Powell (1987). It is interesting to note that the cross sectional estimator that uses all pairs of observations is $\sqrt{n}$ consistent in both of these cases, although the corresponding panel data estimator converges at a slower rate.

The situation is a little more complicated for the monotone transformation model because the panel data estimator of that model is itself based on pairwise comparisons across individuals. The cross sectional version that treats each pair of observations as if they came from a panel of length 2, is therefore based on comparing pairs of pairs, resulting in an estimator defined by a quadruple sum. This estimator is analyzed in Abrevaya (1999).

Table 3 summarizes the relationship between the panel data estimators and their pairwise comparison counterparts. It also lists the estimator for the cross sectional model which we find to be closest in spirit to the panel data estimator.

Table 3
Relationship between panel data estimators and pairwise comparison estimators

Model	'Motivating' estimator	Panel data estimator	Pairwise comparison
Discrete choice	Manski (1975)	Manski (1987)	Han (1987)
Censored regression	Powell (1986)	Honoré (1992)	Honoré and Powell (1994)
Selection	Powell (1987)	Kyriazidou (1997)	Powell (1987)
Type 3 Tobit		Honoré and Kyriazidou (2000b)	Honoré et al. (1997)
Monotone transformation	Han (1987)	Abrevaya (1999)	Abrevaya (1999)

8. Models with lagged dependent variables

With the exception of the models with multiplicative effects, the non-linear models discussed so far all assume that the explanatory variables are strictly exogenous. This assumption is in sharp contrast to the discussion in the first part of this chapter which focused on linear models with predetermined variables. The assumption of strict exogeneity is important. For example, with two time-periods, the basic idea in the logit model was to consider the probability that $y_{i1} = 1$ conditional on the explanatory variables in both periods and conditional on $y_{i1} \neq y_{i2}$. If the explanatory variables include a lagged dependent variable, then the conditioning set includes y_{i1} and $y_{i1} \neq y_{i2}$. This means that the probability is either 1 or zero and cannot be used to make inference about β. By reviewing each of the other methods described in the previous section, it is clear that the motivation for all of them is based on some statement about the joint distribution of (y_{i1}, y_{i2}) given (x_{i1}, x_{i2}). If the explanatory variable in the second time-period, x_{i2}, includes the lagged dependent variable, y_{i1}, then the arguments fail.

In this section, we will review some recently proposed methods for dealing with lagged dependent variables in nonlinear models with fixed effects. It will be seen that some progress has been made in this area, but that the methods that have been proposed are case-specific and often lead to estimators that do not converge at the usual $\sqrt{n}$ rate. One might conclude from this that it would be more fruitful to take a random effect approach that makes some assumptions on the distribution of the individual-specific effects. However, estimation of dynamic nonlinear models is very difficult even in that case. The main difficulty is the so-called initial conditions problem: if one starts observing the individuals when the process in question is already in progress, then the first observation will depend on the dependent variable in the period before the sample starts. Even if that is observed (or one drops the first observation) one will have to deal with relationship between the first lagged dependent variable and the individual-specific effect. That relationship will depend (in a complicated way) on the parameters of the model, but also on the distribution of the explanatory variables in periods prior to the start of the sample, which is typically unknown. In practice one might “solve” this problem by assuming a flexible functional form for the distribution of the first observation (see for example Heckman (1981b) for a discussion of this approach). One case where one can ignore the initial conditions problem is when one can reasonably assume that the process is observed from the start. For example, if the dependent variable is labor supply and the sample consists of people observed (say) from the time they graduated from high school, then there will be no initial conditions problem.

In the next three sections we discuss some approaches that have been used to generalize the limited dependent variable models discussed earlier to the case where one of the explanatory variables is the lagged dependent variable. Very little is known about how to deal with general predetermined variables in the models that we consider.

8.1. Discrete choice with state dependence

Including a lagged dependent variable among the explanatory variables in the discrete choice model with individual specific effects gives the model

$$y_{it} = 1\left\{x_{it}\beta + \gamma y_{i,t-1} + \alpha_i + \varepsilon_{it} \geqslant 0\right\} \quad t = 1, \ldots, T;\ i = 1, \ldots, n. \tag{125}$$

In its most general setting, this model allows for three sources of persistence (after controlling for the observed explanatory variable x) in the event described by y_{it}. Persistence can be the result of serial correlation in the error term ε, a result of the “unobserved hererogeneity” α, or a result of true state dependence through the term $\gamma y_{i,t-1}$. Distinguishing between these sources of persistence is important in many situations because they have very different policy implications. A policy that temporarily increases the probabality that $y = 1$ will have different implications about future probabilities in a model with true state dependence than in model where the persistence is due to unobserved heterogeneity. See, for example, Heckman (1981a) for a discussion of this. Distinguishing between persistence due to state dependence and due to heterogeneity is also important because they sometimes correspond to different economic models. For example, Chiappori and Salanie (2000) and Chiappori (1998) argue that it can be used to distinguish between moral hazard and adverse selection. The pricing system in the French automobile insurance market is such that the incentives for not having an accident are stronger if the driver has had fewer accidents in the past. This suggests that accident data should show true state dependence: having an accident this period should lower the probability of an accident next period. On the other hand adverse selection suggests that some drivers are permanantly more likely to have accidents, which corresponds to the individual specific effect α_i in Equation (125).

It is clear that even if the errors are serially independent, the conditions discussed earlier for conditional maximum likelihood estimation of the fixed effects logit model are not satisfied because they implied that ε in time period t is independent of the explanatory variables in time period $t-1$, a condition which clearly fails when one of the explanatory variables is the lagged dependent variable. By the same argument, the conditions for the conditional maximum score estimator will not be satisfied in the presence of a lagged dependent variable. On the other hand it is also clear that the two sources of persistence in Equation (125) have very different implications. For example consider the case where there are no other explanatory variables: if there is no "state dependence" ($\gamma = 0$) then the sequence $(0, 1, 0, 1)$ would be as likely as the sequence $(0, 0, 1, 1)$. On the other hand, if $\gamma < 0$ then the first sequence would be more likely, whereas the second would be more likely if $\gamma > 0$. As pointed out by Heckman (1978), this suggests that one should be able to test for "no state dependence" in a model like Equation (125). As will be seen below, this observation can also be used to estimate γ and β in Equation (125).

Consider first the special case of a logit model where the lagged dependent variable is the *only* explanatory variable,

$$y_{it} = 1\left\{\gamma y_{i,t-1} + \alpha_i + \varepsilon_{it} \geqslant 0\right\} \quad t = 1, \ldots, T;\ i = 1, \ldots, n,$$

where ε_{it} is i.i.d., independent of α_i, and logistically distributed. Considering only the first three observations (and the initial condition), we have

$$\Pr\left(y_{it} = 1 | \alpha_i, y_{i0}, \ldots, y_{i,t-1}\right) = \frac{\exp(\gamma y_{i,t-1} + \alpha_i)}{1 + \exp(\gamma y_{i,t-1} + \alpha_i)} \quad t = 1, 2, 3.$$

It is then an easy exercise to see that

$$\Pr\left(y_{i1} = 0 | y_{i1} + y_{i2} = 1, \alpha_i, y_{i0}, y_{i3}\right) = \frac{1}{1 + \exp\left(\gamma(y_{i0} - y_{i3})\right)},$$

which does not depend on α_i, and which can therefore be used to make inference on γ [Chamberlain (1978)]. More generally, with T observations for each individual, the conditional distribution of $(y_{i1}, \ldots, y_{iT})$ given $y_{i1}, \sum_{t=1}^{T} y_{it}$ and y_{iT} is

$$P\left(y_{i1}, \ldots, y_{iT} | y_{i1}, \sum_{t=1}^{T} y_{it}, y_{iT}, \alpha_i\right) = \frac{\exp\left(\gamma \sum_{t=2}^{T} y_{it} y_{i,t-1}\right)}{\sum_{(d_1, \ldots, d_t) \in B} \exp\left(\gamma \sum_{t=2}^{T} d_t d_{t-1}\right)}, \tag{126}$$

where B is the set of all sequences of zeros and ones that have $\sum_{t=1}^{T} d_{it} = \sum_{t=1}^{T} y_{it}$, $d_{i1} = y_{i1}$ and $d_{it} = y_{iT}$. Magnac (1997) presents similar results for the multinomial logit version of this model. He also presents the conditional likelihood function for models with more than one lag.

Honoré and Kyriazidou (2000a) modify the calculations leading to the conditional maximum likelihood estimator of a fixed effects logit in such a way that it can be applied to Equation (125). Specifically, assume that ε_{it} in Equation (125) are i.i.d.

logistically distributed and that each observation is observed for at least four periods (three periods in which both the exogenous variables and the dependent variable are observed, plus the initial value of y). Unlike the case where the lagged dependent variable is the only explanatory variable, $P\left(y_{i1},\ldots,y_{iT}\,|\,y_{i0},\sum_{t=1}^{T}y_{it},y_{iT},\{x_{it}\}_{t=1}^{T},\alpha_i\right)$ will in general depend on α_i, and the conditional likelihood approach will therefore generally break down. However (considering the case with $T = 3$ for simplicity), Honoré and Kyriazidou (2000a) showed that

$$P\left(y_{i1},\ldots,y_{i3}\,|\,y_{i0},\sum_{t=1}{}^{3}y_{it},y_{i3},\{x_{it}\}_{t=1}^{3},\alpha_i,x_{i2}=x_{i3}\right)$$
$$=\frac{1}{1+\exp((x_{i1}-x_{i2})\beta+\gamma(y_{i0}-y_{i3}))},\tag{127}$$

which does *not* depend on α_i. This suggests estimating β and γ by maximizing a conditional likelihood function based on Equation (127). However, if one of the explanatory variables is continuously distributed, there will typically be no observations for which $x_{i1} = x_{i2}$. This is similar to the situation when one wants to estimate a conditional expectation of one random variable given that another takes a particular value. One remedy in that case is to use a kernel estimator to average over observations close to the value. Based on this idea, Honoré and Kyriazidou (2000a) estimate γ and β by

$$(\hat{\beta},\hat{\gamma})=\arg\max_{(b,g)}\sum_{i=1}^{n}\quad 1\{y_{i1}+y_{i2}=1\}K\left(\frac{x_{i2}-x_{i3}}{h}\right)$$
$$\times\ln\left(\frac{\exp((x_{i1}-x_{i2})\,b+g(d_{i0}-d_{i3}))^{y_{i1}}}{1+\exp((x_{i1}-x_{i2})\,b+g(d_{i0}-d_{i3}))}\right),\tag{128}$$

where $K(\cdot)$ is a kernel[24] which gives the appropriate weight to observation i, and $h \to 0$ as $n \to \infty$. The main limitation of this approach is that it uses only observations in a neighborhood of $x_{i2} = x_{i3}$, so it is necessary to assume that distribution of $x_{i2} - x_{i3}$ to have support in a neighborhood of 0. This rules out time-dummies. Honoré and Kyriazidou (2000a) give conditions under which this estimator is consistent and asymptotically normal (although it does not converge at rate $\sqrt{n}$, and they discuss generalizations to general T, to multinomial models and to models with more lags.

[24] The term $K\left(\frac{x_{i2}-x_{i3}}{h}\right)$ in Equation (128) plays the same role as the kernel does in non-parametric regression. In a sample, there will be no two observations for which $x_i = x_j$ if x is continuously distributed. However if the object of interest (typically the conditional expectation) is sufficiently smooth, then we can use observations where x_i is close to x_j, where "close" is defined appropriately. See, e.g., Härdle and Linton (1994) for a description of non-parametric regression.

The same trick as above can be used to modify Manski's conditional maximum score estimator in such a way that it applies to the model

$$y_{it} = 1\left\{x_{it}\beta + \gamma y_{i,t-1} + \alpha_i + \varepsilon_{it} \geqslant 0\right\} \quad t = 1,2,3;\ i = 1, \ldots, n,$$

where ε_{it} is i.i.d. (independent of (α_i, x_i)) with distribution function F. Specifically,

$$\begin{aligned}
&\operatorname{sgn}\left(P\left(y_{i2} = 1 \mid y_{i0}, \sum_{t=1}^{3} y_{it}, y_{i3}, \{x_{it}\}_{t=1}^{3}, \alpha_i, x_{i2} = x_{i3}\right)\right.\\
&\quad \left.- P\left(y_{i1} = 1 \mid y_{i0}, \sum_{t=1}^{3} y_{it}, y_{i3}, \{x_{it}\}_{t=1}^{3}, \alpha_i, x_{i2} = x_{i3}\right)\right)\\
&= \operatorname{sgn}\left((x_{i2} - x_{i1})\beta + \gamma(d_{i3} - d_{i0})\right).
\end{aligned}$$

Mimicking the logic in Manski (1987), this means that we can consistenty estimate β and γ up to scale by

$$\begin{aligned}
\left(\hat{\beta}, \hat{\gamma}\right) &= \arg\max_{(b,g)} \sum_{i=1}^{n} K\left(\frac{x_{i2} - x_{i3}}{h}\right) \operatorname{sgn}\left(y_{i2} - y_{i1}\right)\\
&\quad \cdot \operatorname{sgn}\left((x_{i2} - x_{i1})b + g(d_{i3} - d_{i0})\right).
\end{aligned}$$

8.2. Dynamic Tobit models

We next turn to the possibility of allowing lagged dependent variables to enter the censored regression model considered earlier. Depending on the context, the relevant lagged dependent variable is either the lagged observed variable or the lagged latent (unobserved variable). Here, we consider only the former case. Specifically, assume that

$$y_{it} = \max\left\{0, \alpha_i + x_{it}\beta + \sum_{\ell=1}^{L} \gamma_\ell y_{i,t-\ell} + \varepsilon_{it}\right\} \qquad t = 1, \ldots, T \quad i = 1, \ldots, n. \tag{129}$$

Honoré (1993) demonstrated that for this model, it is possible to obtain moment conditions that must be satisfied at the true parameter values. To see how this can be done, assume that $\gamma_\ell \geq 0$ for $\ell = 1, \ldots, L$, and define "residuals" by

$$\upsilon_{ist}(b, g) \equiv \max\left\{0, (x_{it} - x_{is})b, y_{it} - \sum_{\ell=1}^{L} g_\ell y_{i,t-\ell}\right\} - x_{it}b.$$

Then

$$\begin{aligned}
\upsilon_{ist}(\beta, \gamma) &\equiv \max\left\{0, (x_{it} - x_{is})\beta, y_{it} - \sum_{\ell=1}^{L} \gamma_\ell y_{i,t-\ell}\right\} - x_{it}\beta\\
&= \max\left\{-x_{it}\beta, -x_{is}\beta, \alpha_i + \varepsilon_{it}\right\}.
\end{aligned}$$

If $\{x_{it}\}_{t=1}^{T}$ is strictly exogenous in the sense that ε_{it} and ε_{is} are identically distributed conditional on $\{x_{it}\}_{t=1}^{T}$ then for any function $\psi(\cdot)$,

$$E\left[\psi\left(v_{ist}(\beta,\gamma)\right)-\psi\left(v_{its}(\beta,\gamma)\right)\middle|\,\{x_{it}\}_{t=1}^{T}\right]=0, \tag{130}$$

which suggests that (β,γ) can be estimated by GMM. Honoré and Hu (2001) present a set of sufficient conditions under which Equation (130) is uniquely satisfied at the true parameter value. The most restrictive assumption is that $x_{it}-x_{is}$ has support in a neighborhood around 0, which rules out time-dummies.

Honoré and Hu (2001) also discuss how a modification of the same idea can be used to construct moment conditions for a model with general predetermined explanatory variables, and Hu (2000) shows how to generalize the approach so that it can be used to construct moment conditions for a model in which the lagged variables in Equation (129) are the lagged uncensored variables. This is, for example, the relevant model if the censoring is due to top-coding.

8.3. Dynamic sample selection models

Kyriazidou (1999) generalizes her approach to estimation of

$$\begin{aligned}
y_{it}^{*} &= \rho_0 y_{it-1}^{*}+x_{it}^{*}\beta_0+\alpha_i^{*}+\varepsilon_{it}^{*}\\
y_{it} &= d_{it}\,y_{it}^{*}\\
d_{it} &= 1\left\{\phi_0 d_{it-1}+w_{it}\gamma_0+\eta_i-u_{it}\leqslant 0\right\}.
\end{aligned}$$

This is the same model that was considered in Kyriazidou (1997), except that the model is now dynamic, with both the dependent variables, y_{it}^{*} and d_{it}, depending on their own lagged value. The key insight is to combine the insights from the dynamic linear panel data models with the insight in Kyriazidou (1997). For simplicity assume that $\left(\varepsilon_{it}^{*},u_{it}\right)$ is i.i.d. over time and independent of all other right hand side variables. Applying the methods discussed in the first part of this chapter to observations for which y_{it}^{*} is observed in three consequtive periods (so $d_{it}=d_{it-1}=d_{it-2}=1$), will result in a sample selection bias term which after first differencing has the form $E\left[\varepsilon_{it}^{*}\middle|\,u_{it}\geqslant\phi_0+w_{it}\gamma_0+\eta_i\right]-E\left[\varepsilon_{it-1}^{*}\middle|\,u_{it-1}\geqslant\phi_0+w_{it-1}\gamma_0+\eta_i\right]$. This sample selection term will be 0 for observations for whom $w_{it}\gamma_0=w_{it-1}\gamma_0$. The idea therefore is to apply the methods discussed in the first part of this paper augmented by kernel-weights that give more weight to observations for which $w_{it}\widehat{\gamma}$ is close to $w_{it-1}\widehat{\gamma}$, where $\widehat{\gamma}$ is an estimate of γ_0 [using, for example, the method proposed in Honoré and Kyriazidou (2000a)].

9. "Random" effects models

Since little is known about how to deal with fixed effects in nonlinear models other than the ones discussed above, it is often appealing to make assumptions on the distribution of the individual effects. When the distribution of the error is parameterized completely, then the resulting model is usually refered to as random effects model. As mentioned in the previous section, this approach is problematic in dynamic models if one does not observe the start of the process. On the other hand, there are no conceptual difficulties in estimating the parameters of a random effects model by maximum likelihood or methods of moments if the explanatory variables are strictly exogenous, and the distribution of the errors, ε_{it}, is specified. The downside is that there might be practical difficulties in implementing these methods, since the likelihood function and the conditional moments will typically involve multivariate integration. In that case, simulation based inference can be extremely useful. See for example Hajivassiliou and Ruud (1994) or Keane (1994). It is also straightforward to consistently estimate the parameters of certain semiparametric random effects models. Consider for example the censored regression model in Section 7.1. If the errors and the individual specific effects are independent of each other and both are independent of the regressors, then β can be estimated by applying one of the many semiparametric estimators of the censored regression model to the pooled data set consisting of the observations for all i and t. The main complication in that case is that one must correct the variance of the estimator to account for the fact that the observations for a given i are not independent (because they all depend on the same individual-specific effect).

A number of papers propose estimators of models that make assumptions that fall between fixed and random effects models. These papers are motivated by the tradeoff between the difficulties in estimating fixed effects versions of nonlinear models and the fairly strong assumptions that one must make in a random effects approach. As an example, consider the discrete choice model of Section 6. Following Chamberlain (1984), if the individual specific effect, α_i, happens to be of the form $\alpha_i = \sum_{t=1}^{T} x'_{it}\gamma_t + u_i$ where u_i and the transitory errors, ε_{it}, are jointly independent of $(x_{i1}, \ldots, x_{iT})$ then one can apply an estimator of the semiparametric discrete choice model to the data for each time-period to estimate $(\gamma_1, \gamma_2, \ldots, \gamma_{t-1}, \gamma_t + \beta, \gamma_{t+1}, \ldots, \gamma_T)$ up to scale. These can then be combined (via minimum distance) to obtain estimators of $\{\gamma_t\}_{t=1}^{T}$ and β (up to scale). In Chamberlain's example, the ε_{it}'s and the u_i's were assumed to be normally distributed, so the estimation could be done by probit maximum likelihood. Although the functional form assumption made on the individual specific effect makes the model much less general than the fixed effects model, it should be noted that the approach does not require the transitory errors to be homoskedastic over time. This is in contrast to the fixed effects estimators which all assumed some kind of stationarity of the errors.

Newey (1994) considered estimation of Chamberlain's model but with $\alpha_i = \rho(x_{i1}, \ldots, x_{iT}) + u_i$ where the function ρ is unknown. If F_t is the cumulative distribution function for $u_i + \varepsilon_{it}$ then

$$P(y_{it} = 1 | x_{i1}, \ldots, x_{iT}) = F_t(\rho(x_{i1}, \ldots, x_{iT}) + x_{it}\beta),$$

or

$$F_t^{-1}(P(y_{it} = 1 | x_{i1}, \ldots, x_{iT})) = \rho(x_{i1}, \ldots, x_{iT}) + x_{it}\beta. \tag{131}$$

When the errors are jointly normally distributed, this implies

$$\begin{aligned}\Phi^{-1}\left(P\left(y_{it}=1\,|\,x_{i1},\ldots,x_{iT}\right)\right) &= \sqrt{\frac{\operatorname{Var}\left[u_i+\varepsilon_{is}\right]}{V\left[u_i+\varepsilon_{it}\right]}}\,\Phi^{-1}\left(P\left(y_{is}=1\,|\,x_{i1},\ldots,x_{iT}\right)\right)\\ &\quad+\sqrt{\frac{1}{\operatorname{Var}\left[u_i+\varepsilon_{it}\right]}}\left(x_{it}-x_{is}\right)\beta.\end{aligned}$$

Since discrete choice models can only be estimated up to scale, one can normalize $\operatorname{Var}[u_i+\varepsilon_{it}] = 1$ and then estimate β and $\sqrt{\operatorname{Var}[u_i+\varepsilon_{is}]}$ by regressing a nonparametric estimate of $P(y_{it}=1\,|\,x_{i1},\ldots,x_{iT})$ on a nonparametric estimate of $P(y_{is}=1\,|\,x_{i1},\ldots,x_{iT})$ and on $(x_{it}-x_{is})$. Newey (1994) derived the limiting distribution of this estimator. Chen (1998) generalized the model further by allowing the distribution of the errors u and ε to be unknown. His insight is to note that if one normalizes one of the components (say, the first) of β to be one so $\beta = \begin{pmatrix}1\\ \widetilde{\beta}\end{pmatrix}$ then Equation (131) implies that

$$x_{it}^{1} = -\rho\left(x_{i1},\ldots,x_{iT}\right) - \tilde{x}_{it}\widetilde{\beta} + F_t^{-1}\left(P\left(y_{it}=1\,|\,x_{i1},\ldots,x_{iT}\right)\right),$$

or

$$x_{it}^{1} - x_{is}^{1} = -\left(\tilde{x}_{it}-\tilde{x}_{is}\right)\widetilde{\beta} + F_t^{-1}\left(P\left(y_{it}=1\,|\,x_{i1},\ldots,x_{iT}\right)\right) - F_s^{-1}\left(P\left(y_{is}=1\,|\,x_{i1},\ldots,x_{iT}\right)\right). \tag{132}$$

Here $P(y_{it}=1\,|\,x_{i1},\ldots,x_{iT})$ and $P(y_{is}=1\,|\,x_{i1},\ldots,x_{iT})$ can be estimated nonparametrically and $\widetilde{\beta}$ can be estimated by observing that Equation (132) is a partially linear regression model of the type studied by e.g., Robinson (1988).

The idea of writing the individual specific effect as $\alpha_i = \rho(x_{i1},\ldots,x_{iT}) + u_i$ where u_i is treated as an error term can also be applied to the other models discussed above. See for example Jacubson (1988) or Charlier, Melenberg and van Soest (2000) for applications of this idea in the context of the censored regression model, and Nijman and Verbeek (1992), Zabel (1992) and Wooldridge (1995) for a discussion of this approach in sample selection models.

In a linear model, there is no loss of generality in making assumptions of the form $\alpha_i = \sum_{t=1}^{T} x_{it}'\gamma_t + u_i$ because one can always interpret $\sum_{t=1}^{T} x_{it}'\gamma_t$ as the projection of α_i on $(x_{i1},\ldots,x_{iT})$. Making such an assumption in a non-linear model is much more restrictive. In particular, if $\alpha_i = \rho(x_{i1},\ldots,x_{iT}) + u_i$ where u_i is independent of $(x_{i1},\ldots,x_{iT})$ for some T then the same assumption will typically not be satisfied for some other T. This means that the model which is estimated (and which is assumed to be true) depends on the number of time-series observations the econometrician happens to have.

Other alternatives to the "pure" fixed approach have been proposed. For example, Lee (1999) makes assumptions on the joint distribution of the regressors and the individual specific effects which allow him to construct a maximum rank correlation-type estimator of the static discrete choice panel data model. Honoré and Lewbel (2000) exploit the assumption that one of the regressors is independent of the individual specific effect to construct an estimator of a discrete choice panel data model with predetermined explanatory variables.

10. Concluding remarks

Our discussion has focused on two of the developments in panel data econometrics since the Handbook chapter by Chamberlain (1984). In the first part of the paper we have reviewed linear panel data models with predetermined variables, and in the second we have discussed methods for dealing with nonlinear panel data models. Unfortunately, the intersection of these two literatures is very small. With the exception of multiplicative models and models where the only source of "predeterminedness" is lagged dependent variables, almost nothing is known about nonlinear models with general predetermined variables. One step in this direction was taken by Arellano and Carrasco (1996). This is an exciting area for future research.

References

Abowd, J.M., and D. Card (1989), "On the covariance structure of earnings and hours changes", Econometrica 57:411–445.

Abrevaya, J. (1999), "Leapfrog estimation of a fixed-effects model with unknown transformation of the dependent variable", Journal of Econometrics 93(2):203–228.

Abrevaya, J. (2000), "Rank estimation of a generalized fixed-effects regression model", Journal of Econometrics 95(1):1–23.

Ahn, S., and P. Schmidt (1995), "Efficient estimation of models for dynamic panel data", Journal of Econometrics 68:5–27.

Ai, C., and C. Chen (1992), "Estimation of a fixed effect bivariate censored regression model", Economic Letters 40:403–406.

Allenby, G.M., and P.E. Rossi (1999), "Marketing models of consumer heterogeneity", Journal of Econometrics 89:57–78.

Alonso-Borrego, C., and M. Arellano (1999), "Symmetrically normalized instrumental-variable estimation using panel data", Journal of Business & Economic Statistics 17:36–49.

Alvarez, J., and M. Arellano (1998), "The time series and cross-section asymptotics of dynamic panel data estimators", Working Paper 9808 (CEMFI, Madrid).

Amemiya, T. (1985), Advanced Econometrics (Harvard University Press).

Amemiya, T., and T.E. MaCurdy (1986), "Instrumental-variable estimation of an error-components model", Econometrica 54:869–881.

Andersen, E.B. (1970), "Asymptotic properties of conditional maximum likelihood estimators", Journal of the Royal Statistical Society, Series B 32:283–301.

Anderson, T.W., and C. Hsiao (1981), "Estimation of dynamic models with error components", Journal of the American Statistical Association 76:598–606.

Anderson, T.W., and C. Hsiao (1982), "Formulation and estimation of dynamic models using panel data", Journal of Econometrics 18:47–82.

Anderson, T.W., N. Kunitomo and T. Sawa (1982), "Evaluation of the distribution function of the limited information maximum likelihood estimator", Econometrica 50(4):1009–1027.

Arellano, M. (1990), "Testing for autocorrelation in dynamic random effects models", Review of Economic Studies 57:127–134.

Arellano, M. (1993), "On the testing of correlated effects with panel data", Journal of Econometrics 59:87–97.

Arellano, M., and S.R. Bond (1988), "Dynamic panel data estimation using DPD – A guide for users", Working Paper 88/15, (Institute for Fiscal Studies, London).

Arellano, M., and S.R. Bond (1991), "Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations", Review of Economic Studies 58:277–297.

Arellano, M., and O. Bover (1995), "Another look at the instrumental-variable estimation of error-components models", Journal of Econometrics 68:29–51.

Arellano, M., and R. Carrasco (1996), "Binary choice panel data models with predetermined variables", Working Paper 9618 (CEMFI, Madrid).

Ashenfelter, O., and D. Card (1985), "Using the longitudinal structure of earnings to estimate the effect of training programs", Review of Economics and Statistics 67:648–660.

Back, K., and D.P. Brown (1993), "Implied probabilities in GMM estimators", Econometrica 61:971–975.

Baltagi, B. (1995), Econometric Analysis of Panel Data (John Wiley and Sons Ltd.)

Baltagi, B., J. Hidalgo and Q. Li (1996), "A nonparametric poolability test", Journal of Econometrics 75:345–367.

Becker, G., M. Grossman and K. Murphy (1994), "An empirical analysis of cigarette addiction", American Economic Review 84:396–418.

Bekker, P.A. (1994), "Alternative approximations to the distributions of instrumental variable estimators", Econometrica 62:657–681.

Bhargava, A., and J.D. Sargan (1983), "Estimating dynamic random effects models from panel data covering short time periods", Econometrica 51:1635–1659.

Blundell, R., and S. Bond (1998), "Initial conditions and moment restrictions in dynamic panel data models", Journal of Econometrics 87:115–143.

Blundell, R., and S. Bond (1999), "GMM estimation with persistent panel data: an application to production functions", Working Paper W99/4 (Institute for Fiscal Studies, London).

Blundell, R., S. Bond, M.P. Devereux and F. Schiantarelli (1992), "Investment and Tobin's Q: evidence from company panel data", Journal of Econometrics 51:233–257.

Blundell, R., R. Griffith and F. Windmeijer (1997), "Individual effects and dynamic count data", Unpublished manuscript (University College London).

Bond, S., and C. Meghir (1994), "Dynamic investment models and the firm's financial policy", Review of Economic Studies 61:197–222.

Bover, O. (1991), "Relaxing intertemporal separability: a rational habits model of labor supply estimated from panel data", Journal of Labor Economics 9:85–100.

Breusch, T.S., G.E. Mizon and P. Schmidt (1989), "Efficient estimation using panel data", Econometrica 57:695–700.

Browning, M. (1992), "Children and household economic behaviour", Journal of Economic Literature 30:1434–1475.

Canova, F., and A. Marcet (1995), "The poor stay poor: nonconvergence across countries and regions", Economics Working Paper 137 (Universitat Pompeu Fabra, Barcelona).

Carrasco, R. (1998), "Binary choice with binary endogenous regressors in panel data: estimating the effect of fertility on female labour participation", Working Paper 9805 (CEMFI, Madrid).

Cavanagh, C.L. (1987), "The limiting behavior of estimators defined by optimization", Unpublished manuscript (Department of Economics, Harvard University).

Chamberlain, G. (1978), "On the use of panel data", Unpublished manuscript (Department of Economics, Harvard University).

Chamberlain, G. (1980), “Analysis of covariance with qualitative data”, Review of Economic Studies 47:225–238.

Chamberlain, G. (1982a), “Multivariate regression models for panel data”, Journal of Econometrics 18:5–46.

Chamberlain, G. (1982b), “The general equivalence of Granger and Sims causality”, Econometrica 50:569–581.

Chamberlain, G. (1984), “Panel data”, in: Z. Griliches and M.D. Intriligator, eds., Handbook of Econometrics, Vol. 2 (Elsevier Science, Amsterdam).

Chamberlain, G. (1987), “Asymptotic efficiency in estimation with conditional moment restrictions”, Journal of Econometrics 34:305–334.

Chamberlain, G. (1992a), “Efficiency bounds for semiparametric regression”, Econometrica 60:567–596.

Chamberlain, G. (1992b), “Comment: sequential moment restrictions in panel data”, Journal of Business & Economic Statistics 10:20–26.

Chamberlain, G. (1993), “Feedback in panel data models”, Unpublished manuscript (Department of Economics, Harvard University).

Chamberlain, G., and K. Hirano (1999), “Predictive distributions based on longitudinal earnings data”, Annales d’Économie et de Statistique 55–56:211–242.

Charlier, E., B. Melenberg and A. van Soest (1995), “A smoothed maximum score estimator for the binary choice panel data model and an application to labour force participation”, Statistica Neerlandica 49:324–342.

Charlier, E., B. Melenberg and A. van Soest (2000), “Estimation of a censored regression panel data model using conditional moments restrictions efficiently”, Journal of Econometrics 95(1):25–56.

Chen, S. (1998), “Root–N consistent estimation of a panel data sample selection model”, unpublished manuscript (The Hong Kong University of Science and Technology).

Chen, S., J.J. Heckman and E. Vytlacil (1998), “Identification and $\sqrt{n}$ estimation of semiparametric panel data models with binary variables and latent factors”, Unpublished manuscript (Department of Economics, University of Chicago).

Chiappori, P.-A. (1998), “Econometric models of insurance under asymmetric information”, Unpublished manuscript (Department of Economics, University of Chicago).

Chiappori, P.-A., and B. Salanie (2000), “Testing for adverse selection in insurance markets”, Journal of Political Economy 108:56–78.

Collado, M.D. (1997), “Estimating dynamic models from time series of independent cross-sections”, Journal of Econometrics 82:37–62.

Crepon, B., F. Kramarz and A. Trognon (1997), “Parameters of interest, nuisance parameters and orthogonality conditions. An application to autoregressive error component models”, Journal of Econometrics 82:135–156.

Deaton, A. (1985), “Panel data from time series of cross-sections”, Journal of Econometrics 30:109–126.

Geweke, J., and M. Keane (2000), “An empirical analysis of earnings dynamics among men in the PSID: 1968–1989”, Journal of Econometrics 96:293–356.

Granger, C.W.J. (1969), “Investigating causal relations by econometric models and cross-spectral methods”, Econometrica 37:424–438.

Griliches, Z., and J.A. Hausman (1986), “Errors in variables in panel data”, Journal of Econometrics 31:93–118.

Hahn, J. (1997), “Efficient estimation of panel data models with sequential moment restrictions”, Journal of Econometrics 79:1–21.

Hahn, J. (1998), “Asymptotically unbiased inference of dynamic panel data model with fixed effects when both n and T are large”, Unpublished manuscript (University of Pennsylvania).

Hajivassiliou, V., and D. McFadden (1990), “The method of simulated scores for the estimation of LDV models with an application to external debt crisis”, Cowles Foundation Discussion Paper 967 (Cowles Foundation for Research in Economics, Yale University).

Hajivassiliou, V., and P.A. Ruud (1994), “Classical estimation methods for LDV models using simulation”, in: R.F. Engle and D.L. McFadden, eds., Handbook of Econometrics, Vol. 4 (Elsevier, Amsterdam) Ch. 40.

Hall, R.E., and F.S. Mishkin (1982), “The sensitivity of consumption to transitory income: estimates from panel data on households”, Econometrica 50:461–481.

Han, A.K. (1987), “Non-parametric analysis of a generalized regression model: the maximum rank correlation estimator”, Journal of Econometrics 35:303–316.

Hansen, L.P. (1982), “Large sample properties of generalized method of moments estimators”, Econometrica 50:1029–1054.

Hansen, L.P., J. Heaton and A. Yaron (1996), “Finite sample properties of some alternative GMM estimators”, Journal of Business & Economic Statistics 14:262–280.

Härdle, W., and O. Linton (1994), “Applied nonparametric methods”, in: R.F. Engle and D.L. McFadden, eds., Handbook of Econometrics, Vol. 4 (Elsevier, Amsterdam) Ch. 38.

Hausman, J.A., and W.E. Taylor (1981), “Panel data and unobservable individual effects”, Econometrica 49:1377–1398.

Hausman, J.A., B. Hall and Z. Griliches (1984), “Econometric models for count data with an application to the patents-R&D relationship”, Econometrica 52(4):909–938.

Hayashi, F., and T. Inoue (1991), “The relation between firm growth and Q with multiple capital goods: theory and evidence from panel data on Japanese firms”, Econometrica 59:731–753.

Hayashi, F., and C.A. Sims (1983), “Nearly efficient estimation of time series models with predetermined, but not exogenous, instruments”, Econometrica 51:783–798.

Heckman, J.J. (1976), “The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models”, Annals of Economic and Social Measurement 15:475–492.

Heckman, J.J. (1978), “Simple statistical models for discrete panel data developed and applied to tests of the hypothesis of true state dependence against the hypothesis of spurious state dependence”, Annales de l’INSEE 30–31:227–269

Heckman, J.J. (1979), “Sample selection bias as a specification error”, Econometrica 47:153–161.

Heckman, J.J. (1981a), “Statistical models for discrete panel data”, in: C.F. Manski and D. McFadden, eds., Structural Analysis of Discrete Panel Data with Econometric Applications (MIT Press).

Heckman, J.J. (1981b), “The incedental parameters problem and the problem of initial conditions in estimating a discrete time–discrete data stochastic process”, in: C.F. Manski and D. McFadden eds., Structural Analysis of Discrete Panel Data with Econometric Applications (MIT Press).

Heckman, J.J., and T.E. MaCurdy (1980), “A life cycle model of female labour supply”, Review of Economic Studies 47:47–74.

Hillier, G.H. (1990), “On the normalization of structural equations: properties of direction estimators”, Econometrica 58:1181–1194.

Holtz-Eakin, D., W.K. Newey and H. Rosen (1988), “Estimating vector autoregressions with panel data”, Econometrica 56:1371–1395.

Honoré, B.E. (1992), “Trimmed lad and least squares estimation of truncated and censored regression models with fixed effects”, Econometrica 60:533–565.

Honoré, B.E. (1993), “Orthogonality conditions for tobit models with fixed effects and lagged dependent variables”, Journal of Econometrics 59:35–61.

Honoré, B.E., and L. Hu (2001), “Estimation of censored regression models with endogeneity”, Unpublished manuscript (Department of Economics, Princeton University).

Honoré, B.E., and E. Kyriazidou (2000a), “Panel data discrete choice models with lagged dependent variables”, Econometrica 68(4):839–874.

Honoré, B.E., and E. Kyriazidou (2000b), “Estimation of Tobit-type models with individual specific effects”, Econometric Reviews 19(3):341–366.

Honoré, B.E., and A. Lewbel (2000), “Semiparametric binary choice panel data models without strictly exogeneous regressors”, Unpublished manuscript (Department of Economics, Princeton University).

Honoré, B.E., and J.L. Powell (1994), "Pairwise difference estimators of censored and truncated regression models", Journal of Econometrics 64(2):241–278.

Honoré, B.E., E. Kyriazidou and C. Udry (1997), "Estimation of type 3 Tobit models using symmetric trimming and pairwise comparisons", Journal of Econometrics 76:107–128.

Horowitz, J.L. (1992), "A smoothed maximum score estimator for the binary response model", Econometrica 60:505–531.

Horowitz, J.L., and M. Markatou (1996), "Semiparametric estimation of regression models for panel data", Review of Economic Studies 63:145–168.

Hsiao, C. (1986), Econometric Analysis of Panel Data (Cambridge University Press).

Hu, L. (2000), "Estimating a censored dynamic panel data model with an application to earnings dynamics", Unpublished manuscript (Department of Economics, Northwestern University).

Imbens, G. (1997), "One-step estimators for over-identified generalized method of moments models", Review of Economic Studies 64:359–383.

Imbens, G., R. Spady and P. Johnson (1998), "Information theoretic approaches to inference in moment condition models", Econometrica 66:333–357.

Jacubson, G. (1988), "The sensitivity of labor supply parameter estimates to unobserved individual effects: fixed and random effects estimates in a nonlinear model using panel data", Journal of Labor Economics 6:302–329.

Kao, C. (1999), "Spurious regression and residual-based tests for cointegration in panel data", Journal of Econometrics 90:1–44.

Keane, M. (1993), "Simulation estimation for panel data models with limited dependent variables", in: G.S. Maddala, C.R. Rao and H.D. Vinod, eds., Handbook of Statistics, Vol. 11 (Elsevier Science, Amsterdam).

Keane, M. (1994), "A computationally practical simulation estimator for panel data", Econometrica 62:95–116.

Keane, M., and D.E. Runkle (1992), "On the estimation of panel-data models with serial correlation when instruments are not strictly exogenous", Journal of Business & Economic Statistics 10:1–9.

Kim, J., and D. Pollard (1990), "Cube root asymptotics", Annals of Statistics 18:191–219.

Kiviet, J.F. (1995), "On bias, inconsistency, and efficiency of various estimators in dynamic panel data models", Journal of Econometrics 68:53–78.

Kunitomo, N. (1980), "Asymptotic expansions of the distribution of estimators in a linear functional relationship and simultaneous equations", Journal of the American Statistical Society 75:693–700.

Kyriazidou, E. (1995), "Essays in estimation and testing of econometric models", Ph.D. dissertation (Northwestern University).

Kyriazidou, E. (1997), "Estimation of a panel data sample selection model", Econometrica 65:1335–1364.

Kyriazidou, E. (1999), "Estimation of dynamic panel data sample selection models", Unpublished manuscript (Department of Economics, University of Chicago).

Lancaster, T. (1997), "Orthogonal parameters in panel data", Working Paper No. 97-12 (Brown University, Department of Economics).

Lee, M.-J. (1999), "A root-*n* consistent semiparametric estimator for related effect binary response panel data", Econometrica 67:427–434.

Li, Q., and C. Hsiao (1998), "Testing serial correlation in semiparametric panel data models", Journal of Econometrics 87:207–237.

Li, Q., and T. Stengos (1996), "Semiparametric estimation of partially linear panel data models", Journal of Econometrics 71:389–397.

Lillard, L.A., and R.J. Willis (1978), "Dynamic aspects of earnings mobility", Econometrica 46: 985–1012.

MaCurdy, T.E. (1982), "The use of time series processes to model the error structure of earnings in a longitudinal data analysis", Journal of Econometrics 18:83–114.

Magnac, T. (1997), "State dependence and heterogeneity in youth employment histories", Working Paper (INRA and CREST, Paris).

Manski, C.F. (1975), "The maximum score estimation of the stochastic utility model of choice", Journal of Econometrics 3:205–228.

Manski, C.F. (1985), "Semiparametric analysis of discrete response: asymptotic properties of the maximum score estimator", Journal of Econometrics 27:313–333.

Manski, C.F. (1987), "Semiparametric analysis of random effects linear models from binary panel data", Econometrica 55:357–362.

Moffitt, R. (1993), "Identification and estimation of dynamic models with a time series of repeated cross-sections", Journal of Econometrics 59:99–123.

Morimune, K. (1983), "Approximate distributions of k-class estimators when the degree of overidentifiability is large compared with the sample size", Econometrica 51:821–841.

Newey, W.K. (1994), "The asymptotic variance of semiparametric estimators", Econometrica 62:1349–1382.

Neyman, J., and E.L. Scott (1948), "Consistent estimates based on partially consistent observations", Econometrica 16:1–32.

Nickell, S.J. (1981), "Biases in dynamic models with fixed effects", Econometrica 49:1417–1426.

Nijman, T., and M. Verbeek (1992), "Nonresponse in panel data: the impact on estimates of a life cycle consumption function", Journal of Applied Econometrics 7:243–257.

Pesaran, M.H., and R. Smith (1995), "Estimating long-run relationships from dynamic heterogeneous panels", Journal of Econometrics 68:79–113.

Phillips, P.C.B. (1983), "Exact small sample theory in the simultaneous equations model", in: Z. Griliches and M.D. Intriligator, eds., Handbook of Econometrics, Vol. 1 (North-Holland, Amsterdam) Ch. 8.

Phillips, P.C.B., and H.R. Moon (1999), "Linear regression limit theory for nonstationary panel data", Econometrica 67:1057–1111.

Porter, J. (1997), "Nonparametric regression estimation for a panel data model with additive individual effects", Unpublished (Harvard University).

Powell, J.L. (1986), "Symmetrically trimmed least squares estimation for Tobit models", Econometrica 54:1435–1460.

Powell, J.L. (1987), "Semiparametric estimation of bivariate latent models", Working Paper no. 8704 (Social Systems Research Institute, University of Wisconsin-Madison).

Powell, J.L. (1994), "Estimation of semiparametric models", in: R.F. Engle and D.L. McFadden, eds., Handbook of Econometrics, Vol. 4 (Elsevier, Amsterdam) Ch. 41.

Qin, J., and J. Lawless (1994), "Empirical likelihood and general estimating equations", Annals of Statistics 22:300–325.

Rasch, G. (1960), Probabilistic Models for Some Intelligence and Attainment Tests (Denmarks Pædagogiske Institut, Copenhagen).

Rasch, G. (1961), "On the general laws and the meaning of measurement in psychology", Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Vol 4 (University of California Press, Berkeley and Los Angeles).

Robinson, P.M. (1988), "Root-n-consistent semiparametric regression", Econometrica 56:931–954.

Runkle, D.E. (1991), "Liquidity constraints and the permanent income hypothesis: evidence from panel data", Journal of Monetary Economics 97:73–98.

Sargan, J.D. (1958), "The estimation of economic relationships using instrumental variables", Econometrica 26:393–415.

Schmidt, P., S.C. Ahn and D. Wyhowski (1992), "Comment", Journal of Business & Economic Statistics 10:10–14.

Sims, C.A. (1972), "Money, income, and causality", American Economic Review 62:540–552.

Wooldridge, J.M. (1995), "Selection corrections for panel data models under conditional mean independence assumptions", Journal of Econometrics 68:115–132.

Wooldridge, J.M. (1997), "Multiplicative panel data models without the strict exogeneity assumption", Econometric Theory 13:667–678.

Zabel, J.E. (1992), “Estimating fixed and random effects models with selectivity”, Economic Letters 40:269–272.

Zeldes, S.P. (1989), “Consumption and liquidity constraints: an empirical investigation”, Journal of Political Economy 97:305–346.

Ziliak, J.P. (1997), “Efficient estimation with panel data when instruments are predetermined: an empirical comparison of moment-condition estimators”, Journal of Business & Economic Statistics 15:419–431.

Chapter 54

INTERACTIONS-BASED MODELS

WILLIAM A. BROCK[*] and STEVEN N. DURLAUF[**]

Department of Economics, University of Wisconsin, 1180 Observatory Drive,
Madison, WI 53706-1393

Contents

Abstract 3299
Keywords 3299
1. Introduction 3300
  1.1. Neighborhoods and inequality 3301
  1.2. Spatial agglomeration 3302
  1.3. Technology choice 3302
  1.4. Preferences 3302
  1.5. Behavior of political parties 3303
  1.6. Social pathologies 3303
  1.7. Information cascades 3304
  1.8. Evolution of science 3304
  1.9. Chapter objectives 3304
2. Binary choice with social interactions 3305
  2.1. General framework 3305
  2.2. Global interactions 3309
  2.3. Local interactions 3312
  2.4. Relationship to statistical mechanics 3314
  2.5. Social planning problem 3315
  2.6. Linear-in-means model 3318
3. Identification: basic issues 3318
  3.1. Binary choice 3320
  3.2. Linear-in-means model 3322
  3.3. Instruments for unobservable expectations 3324
  3.4. Identification of individual versus neighborhood contextual effects 3325
  3.5. Nonlinear-in-means model 3326
  3.6. Implications of self-selection for identification 3328
  3.7. Implications of multiple equilibria for identification 3331

[*] e-mail: wbrock@ssc.wisc.edu
[**] e-mail: sdurlauf@ssc.wisc.edu

Handbook of Econometrics, Volume 5, Edited by J.J. Heckman and E. Leamer

3.8. Dynamic models and rational expectations 3331
4. Further topics in identification 3335
4.1. Panel data 3335
4.1.1. Fixed effects 3336
4.1.2. Learning 3337
4.2. Duration data 3337
4.3. Nonparametric approaches 3340
4.3.1. Treatment effects 3340
4.3.2. Duration data 3346
5. Sampling properties 3348
5.1. Laws of large numbers 3348
5.2. Naive estimator 3349
5.3. Asymptotics for data generated by social planner 3349
5.4. Unobserved variables 3350
6. Statistical analysis with grouped data 3352
6.1. Differences in cross-group behavior 3353
6.2. Spatial patterns 3353
6.3. Ecological inference 3354
7. Evidence 3355
7.1. Analyses under assumption of correct specification 3355
7.1.1. Neighborhood effects in youth and adult outcomes 3355
7.2. Analyses which are robust to unobserved correlated heterogeneity 3356
7.2.1. Matching 3356
7.2.1.1. Gautreaux Assisted Housing Program 3357
7.2.1.2. Moving to Opportunity Demonstration 3358
7.2.1.3. Milwaukee School Voucher Program 3359
7.2.1.4. Classroom tracking 3360
7.2.1.5. Siblings 3361
7.2.2. Instrumental variables 3361
7.2.2.1. Neighborhood socioeconomic influences 3361
7.2.2.2. Catholic versus public schools 3362
8. Summary and conclusions 3362
Appendix A 3363
A.1. Properties of binary choices made under a social planner 3363
A.1.1. Basics 3363
A.1.2. Asymptotic moments: proof of Theorem 11 3365
A.1.3. Maximum likelihood theory 3367
A.2. Proof of Theorem 5 3368
A.3. Proof of Theorem 7 3369
References 3371

Abstract

This paper describes a range of methods which have been proposed to study interactions in economic and social contexts. By interactions, we refer to interdependences between individual decisions which are not mediated by markets. These types of models have been employed to understand phenomena ranging from the effect of neighborhoods on the life prospects of children to the evolution of political party platforms. We provide a general choice-based framework for modelling such interactions which subsumes a number of specific models which have been studied. This framework illustrates the relationship between interactions-based models and models in statistical mechanics. Our analysis is then extended to the econometrics of these models, with an emphasis on the identification of group-level influences on individual behavior. Finally, we review some of the empirical work on interactions which has appeared in the social science literature.

Keywords

interactions, identification, binary choice, linear-in-means model, nonlinear models, dynamic models, treatment effects

JEL classification: C1, C5, D1, J0

1. Introduction [1]

> "The principal task of the social sciences lies in the explanation of social phenomena, not the behavior of single individuals. In isolated cases the social phenomena may derive directly, through summation, from the behavior of individuals, but more often this is not so. Consequently, the focus must be on the social system whose behavior is to be explained. This may be as small as a dyad or as large as a society or even a world system, but the essential requirement is that the explanatory focus be on the system as a unit, not on the individuals or other components which make it up." (p. 2)
>
> "(A)n internal analysis based on actions and orientations of units at a lower level can be regarded as more fundamental, constituting more nearly a theory of system behavior, than an explanation which remains at the system level...although an explanation which explains the behavior of a social system by the actions and orientations of some entities between the system level and the individual level may be adequate for the purpose at hand, a more fundamental explanation based upon the actions and orientations of individuals is generally more satisfactory." (p. 4)
>
> *James Coleman (1990)*

The role of interactions in economic outcomes has become an important area of research over the last decade. By interactions-based models, we refer to a class of economic environments in which the payoff function of a given agent takes as direct arguments the choices of other agents. The goal of such an analysis is to provide an explanation of group behavior which emerges from the interdependences across individuals.

In some respects, interactions-based models would appear to be nothing but a variant of game-theoretic formulations of decisionmaking; see Blume (1997) and Young (1998) for an excellent syntheses of a number of game-theoretic models from the interactions perspective, and Morris (1998) for a game-theoretic analysis of interaction structures. Further, [Jones (1984), Cooper and John (1988), Milgrom and Roberts (1990)], there has been a great deal of work explicitly focusing on how one type of interaction effects, complementarities, can lead to multiple equilibria and other interesting aggregate phenomena, including breakdowns of the law of large numbers [Jovanovic (1985)]. Indeed, following Bryant (1985), macroeconomic models of complementarities have become a standard research tool. Similarly, analyses such as Bernheim (1994) have shown how conformity effects can produce customs, fads and highly different subcultures within a given population.

[1] This work has benefited from our interactions over the years with Kenneth Arrow, Lawrence Blume, James Heckman, Charles Manski, Peter Phillips and H. Peyton Young. We thank seminar participants at the Brookings Institution, Michigan State University, NBER, Northwestern University, University of Geneva, University of Lausanne, University of Michigan, University of Texas and the University of Wisconsin as well as Kenneth Arrow, Arthur Goldberger, Andros Kourtellos, Artur Minkin, Derek Neal, Eldar Nigmatullin, and the MacArthur Research Network on Social Interactions and Economic Inequality for providing helpful comments on various drafts of this paper. Financial support from the John D. and Catherine T. MacArthur Foundation, National Science Foundation, Romnes Fund, and Vilas Trust is gratefully acknowledged.

Similarly, social sciences other than economics have a much longer tradition of looking for interaction effects. One particularly important example is the Coleman Report of 1966 [Coleman et al. (1966)], which argued that school performance of the disadvantaged was much more amenable to improvement through manipulation of peer group influence than by increased per student expenditures. While the Coleman Report itself has not withstood subsequent scrutiny, its impact on both social science research and public policy was and is immense [see Heckman and Neal (1996) for discussion]. See Blalock (1984) for additional discussion of sociological approaches. Another example is linguistics, where the role of interactions in influencing dialect choice has been well understood for decades [cf. Labov (1972a,b)].

What distinguishes the new research on interactions-based models is the explicit attention given to formulating how each individual's behavior is a function of the characteristics or behavior of others and then studying what aggregate properties emerge in the population. This approach typically, though not always, is done in the form of first specifying a conditional probability measure which describes each individual's behavior as a function of the rest of the population and then determining what joint probability measures are compatible with these conditional measures. This particular approach means that interactions-based approaches have typically been deeply reliant on the use of the probability theory which underlies statistical mechanics methods in physics. (Mathematicians generally refer to statistical mechanics models as interacting particle systems. These models also fall into the broader class of probability models known as random fields.) The value of this approach is that it permits one to specify individual and social aspects of behavior simultaneously, and thereby address aggregate behavior in a way consistent with the sort of methodological individualism advocated by Coleman.

Interactions-based models have been applied to a wide range of contexts both within economics and within social science more generally. A sense of this range can be given through an admittedly incomplete survey of applications; see Durlauf (1997), Kirman (1997), and Rosser (1999) for additional overviews.

1.1. Neighborhoods and inequality

Much of the recent literature on persistent income inequality has focused on the role of neighborhood influences on socioeconomic outcomes. Theoretical models, such as Bénabou (1993, 1996a,b), Cooper (1998), Durlauf (1996a,b), share a common assumption that individual human capital acquisition depends on the behaviors and/or characteristics of community members. These influences may range from peer group effects, in which the costs to one person from investing effort in education are decreasing in the effort levels of others [Bénabou (1993)], to role model effects, in which the aspirations of a student are affected by the observed education/occupation outcomes among adults in his community [Streufert (1991)], to labor market connections [Granovetter (1995), Montgomery (1991, 1992)], in which the probability with which one makes a successful job match depends on the information possessed by members of one's social network. Similar types of spillovers were used much earlier in Loury (1977) to provide a theory of racial income differences. Examples of studies which have adduced empirical evidence of neighborhood effects include Crane (1991a,b), Case and Katz (1991), Haveman and Wolfe (1994). Within the psychology literature, there is rich evidence on the importance of peer group effects, as illustrated, for example, in Brown (1990) and Brown et al. (1986). Finally, recent work by Casella and Rauch (1997, 1998) shows how ethnic social networks can influence patterns of international trade through similar mechanisms with attendant implications for ethnic patterns of inequality.

1.2. Spatial agglomeration

The role of interactions effects in determining location decisions has been analyzed in many contexts. Schelling's (1971) work on racial segregation, illustrates how weak preferences by individuals for neighbors of similar ethnicity can lead to complete segregation. This work is possibly the first interactions-based model to be studied in the social sciences; see Granovetter and Soong (1988) for a number of extensions and generalizations of this original framework. Arthur (1987) has shown how sequential locational decisions, combined with locational spillover effects, can produce agglomerations of economic activity such as the Silicon Valley. Similar models, with a richer microeconomic structure, have been subsequently analyzed by Krugman (1996). In related work, Kelly (1997) has illustrated the evolution of geographically defined trade networks.

1.3. Technology choice

The adoption of particular technological standards is a well-studied case both by economic historians and economic theorists. Standard references on technology adoption and network externalities include Farrell and Saloner (1985) and Katz and Shapiro (1986). David's (1985) discussion of how the QWERTY keyboard became the standard for typewriters is one of the best known examples. Arthur (1989), using mathematical models which fall within the class of tools which are conventionally used in interactions-based models, showed how, when adoption decisions are made sequentially, path dependence in technology choice may occur, which allows inferior technologies to become locked-in. An and Kiefer (1995) show how similar results can occur through local interactions. Goolsbee and Klenow (1998) have provided evidence of the role of interaction effects in home computer adoption.

1.4. Preferences

A number of authors have used interactions-based approaches to study interdependent preferences. Föllmer (1974), in what appears to be the first explicit use of statistical mechanics methods in economics, studied an economy in which the probability that a given individual has one of two utility functions depends on the utility function of his neighbors. His work demonstrated how interactions can lead to breakdowns of the law of large numbers in large economies. Conlisk (1976) showed how to develop Markov chain models in which the distributions of behaviors at $t-1$ determined transition probabilities at t and thereby are capable of producing fads in demand; Granovetter and Soong (1986) developed similar results using different methods. Bell (1995) analyzed a model in which preferences depend on the observed consumption of neighbors. Her work showed how supply effects, in which higher consumption of a commodity by others raises the price of a good for an individual, can be combined with conformity effects, in which higher consumption by others shifts the preferences of an individual toward that commodity, to produce interesting aggregate price dynamics. Darrough et al. (1983), Alessie and Kapteyn (1991), Kapteyn et al. (1997), and Binder and Pesaran (1998b) provide empirical evidence of interaction effects in consumer expenditures using a variety of modelling approaches; Andreoni and Scholz (1998) illustrate similar effects in the context of charitable contributions.

In a complementary line of work, recent authors have considered the implications of concern over relative social position on behavior, an idea whose antecedent is Duesenberry (1949) and which is explored along many dimensions in Frank (1985). Recent important contributions include Cole et al. (1992) who show how relative status concerns can provide a theory of growth, and Clark and Oswald (1996) who show how such concerns affect the relationship between income and well-being, and Clark and Oswald (1998) who characterize the relationship between relative status concerns and emulative behavior. Postlewaite (1997) provides an overview of the relationship between the incorporation of relative status in utility and economic theory.

1.5. Behavior of political parties

Interactions-based methods have recently proven useful in the study of political parties. In a series of papers, Kollman et al. (1992, 1997a,b) have examined the ways in which political parties evolve in response to voter preferences when there are multiple issues of concern. Their modelling typically considers how a political party will adjust its platform in response to the preferences of voters and a consideration of the behavior of the opposing party. This work has illustrated how the convergence of party platforms to a stable configuration depends sensitively on the distribution of voter preferences as well as the degree of foresight of the parties themselves.

1.6. Social pathologies

There exists evidence that a number of types of behavior which society regards as undesirable (pathological) are sustained by interaction effects. One example of this is cigarette smoking. A number of studies [Bauman and Fisher (1986), Krosnick and Judd (1982), Jones (1994)] have directly documented a role for friend and peer group behavior in predicting individual smoking probabilities. Further, well documented differences in smoking rates between black and white teenagers and between men and women within those groupings are highly suggestive of interactions effects. Examples which are closer to the traditional concerns of economists include crime, labor market participation, out-of-wedlock births, and school attendance. Recent theoretical models of interactions and social pathologies include Akerlof and Yellen (1994), Brock and Durlauf (1995), Nechyba (1996), Lindbeck et al. (1999), Sah (1991) and Verbrugge (1999). Statistical evidence of these effects has been found in studies such as Crane (1991a,b), Glaeser et al. (1996), Sampson et al. (1997) and Sucoff and Upchurch (1998); although see Gottfredson and Hirschi (1990) and Sampson and Laub (1995) for skepticism concerning the role of peer group effects with respect to the case of juvenile delinquency. Ethnographic evidence of such interactions may be found in Anderson (1990) and Duneier and Molotch (1999). Finally, Akerlof and Kranton (1998) develop a framework for understanding the psychological bases which lead to memberships in particular reference groups with attendant behavioral implications.

1.7. Information cascades

A number of authors have considered the implications of information aggregation and behavior when agents possess idiosyncratic knowledge and are attempting to learn more by observing the behavior of others. Banerjee (1992) and Bikhchandani et al. (1992) have shown how such behavior can lead to informational cascades and conformity in group behavior. Caplin and Leahy (1994) show how this idea can lead to phenomena such as bank runs; Romer (1993) develops similar results in the context of asset price movements.

1.8. Evolution of science

Since Kuhn's (1970) analysis of scientific paradigms and the nature of scientific revolutions, philosophers of science have grappled with the question of how (and in some cases whether) a community of scientists whose members are subject to conformity effects and whose objectives include non-epistemic factors such as professional status as well as epistemic factors such as better predictability succeeds in shedding scientifically inferior theories for superior ones. Recent work, best exemplified by Kitcher (1993) has explicitly modelled scientific communities as collections of interdependent researchers. This work has led authors such as Dasgupta and David (1994), David (1998), Oomes (1998) and especially Brock and Durlauf (1999) to consider formal interactions models of scientific theory choice. Using interactions-based methods, Brock and Durlauf were able to provide conditions under which scientific evidence will outweigh non-epistemic motivations and thereby provide a model of scientific progress which takes into account critiques of various social constructivists.

1.9. Chapter objectives

This chapter is designed to describe a range of methods to study interactions effects. While the interactions-based models are now fairly well developed from the perspective of theory [see Blume and Durlauf (1998a) for discussion], the econometrics literature is still in its infancy. Most of the existing econometric work has focused on the identification issues which arise for interactions-based models. The pioneering work in this regard is Manski (1993a,b, 1995, 1997); see as well recent surveys by Moffitt (1998) and Duncan and Raudenbusch (1998). Even here, there is substantial work which remains to be done in terms of the analysis of nonlinear as opposed to linear models. A major purpose of this chapter will be to explore identification as well as estimation in the context of structural models of interactions.

In order to facilitate this overview, we will focus on a particular class of interactions-based models, namely binary choice models with interactions. This framework has been exploited by a number of authors, including Blume (1993, 1995), Brock (1993), Brock and Hommes (1998), Durlauf (1993, 1997), and Glaeser et al. (1996). The specific framework we employ is adopted primarily from Brock and Durlauf (1995). Its important advantage, from our perspective, is that for this class of models a tight link exists between the theoretical formulation of various socioeconomic environments and the econometric analysis of those formulations[2].

2. Binary choice with social interactions

2.1. General framework

In this section, we present a baseline model of interactions. The model is capable, for particular restrictions on its parameters, of encompassing many of the theoretical treatments of social interactions which have been developed. An additional purpose of this approach is to show how these models can be analyzed using natural extensions of standard economic reasoning. Finally, as initially recognized by Blume (1993) and Brock (1993), the model is mathematically equivalent to logistic models of discrete choice. This equivalence will allow us to analyze theoretical and econometric aspects of interactions in a common framework.

We consider a population of I individuals each of which faces a binary choice. These choices are denoted by an indicator variable ω_i which has support $\{-1, 1\}$. Each individual makes a choice in order to maximize a payoff function V. In the standard binary choice formulation of economics, this payoff function is of course assumed to depend on the characteristics of the individual in question. These characteristics, in turn, are assumed to be divided into an observable (to the modeller) vector $\boldsymbol{Z}_i$ and a pair of unobservable (to the modeller, but observable to agent i) random shocks $\epsilon_i(1)$, and $\epsilon_i(-1)$. The observable vector can include elements such as family background, role model or peer group characteristics, and past behavior. The shocks $\epsilon_i(1)$ and $\epsilon_i(-1)$ are distinct as various types of unobservable idiosyncrasies are only relevant for one

[2] We will not discuss the branch of the interactions literature which uses computer simulation methods to study various environments. Epstein and Axtell (1996) represent the most ambitious and wide ranging effort yet undertaken in this regard. See also Axtell et al. (1996) for an analysis of how to assess simulations of this type.

of the choices. For example, for the binary choice of whether to remain enrolled or dropout of school, $\epsilon_i(1)$ might refer to a shock which measures unobserved academic ability and so is only relevant if the person stays in school. Algebraically, the individual choices represent the solutions to

$$\max_{\omega_i \in \{-1,1\}} V(\omega_i, \boldsymbol{Z}_i, \epsilon_i(\omega_i)). \tag{1}$$

The standard approach to characterizing the behavior of the population of choices, an approach which renders the model econometrically estimable, is to make some assumption concerning the distribution of the $\epsilon_i(\omega_i)$'s. One common assumption is that the unobservables are independent and extreme value distributed both within and across individuals. This will imply that for a given individual, the difference between the unobservable components is logistically distributed,

$$\mu(\epsilon_i(-1) - \epsilon_i(1) \leqslant z) = \frac{1}{1+\exp(-\beta_i z)}; \qquad \beta_i \geqslant 0. \tag{2}$$

We use $\mu(\cdot)$ to denote probability measures throughout. The subscript i here and elsewhere will be used to capture dependence on $\boldsymbol{Z}_i$ so that, for example, $\beta_i = \beta(\boldsymbol{Z}_i)$.

The interactions-based approach to binary choice, at least qualitatively, is based upon studying this same model once explicit attention has been given to the influence of the expected behavior of others on each individual's choice. Algebraically, each choice is described by

$$\max_{\omega_i \in \{-1,1\}} V(\omega_i, \boldsymbol{Z}_i, \mu_i^e(\boldsymbol{\omega}_{-i}), \epsilon_i(\omega_i)), \tag{3}$$

where $\boldsymbol{\omega}_{-i} = (\omega_1, \ldots, \omega_{i-1}, \ \omega_{i+1}, \ldots, \omega_I)$ denotes the vector of choices other than that of i, and $\mu_i^e(\boldsymbol{\omega}_{-i})$ denotes that individual's beliefs concerning the choices of other agents. The nature of these beliefs, whether they are rational, etc., will be specified below. However, we will assume that beliefs are independent of the realization of any of the $\epsilon_i(\omega_i)$'s.

At this level of generality, there is of course little that can be said about the properties of the population as a whole. Hence, we make two parametric assumptions which will elucidate both basic ideas and will encompass (as special cases) a number of models which have appeared in the literature. First, we assume that the payoff function V can be additively decomposed into three terms.

$$V(\omega_i, \boldsymbol{Z}_i, \mu_i^e(\boldsymbol{\omega}_{-i}), \epsilon_i(\omega_i)) = u(\omega_i, \boldsymbol{Z}_i) + S(\omega_i, \boldsymbol{Z}_i, \mu_i^e(\boldsymbol{\omega}_{-i})) + \epsilon_i(\omega_i). \tag{4}$$

Here $u(\omega_i, \boldsymbol{Z}_i)$, represents deterministic private utility, $S\left(\omega_i, \boldsymbol{Z}_i, \mu_i^e(\boldsymbol{\omega}_{-i})\right)$ represents deterministic social utility, and $\epsilon_i(\omega_i)$ represents random private utility. The two private utility components are standard in the econometric formulations of discrete choice.

The essential difference between recent theoretical work and previous approaches to studying binary choices is the introduction of social utility considerations.

Second, we assume that this social utility term embodies a generalized quadratic conformity effect, i.e.,

$$S\left(\omega_i, \boldsymbol{Z}_i, \mu_i^e\left(\boldsymbol{\omega}_{-i}\right)\right) = -E_i \sum_{j \neq i} \frac{J_{i,j}}{2}\left(\omega_i - \omega_j\right)^2. \tag{5}$$

The term $\frac{J_{i,j}}{2}$ represents the interaction weight which relates i's choice to j's choice and is typically assumed to be nonnegative in theoretical models, although there is no need to do so. We also treat the $J_{i,j}$ parameters as fixed; see Ioannides (1990, 1997a), Kirman (1983), and Kirman et al. (1986) for analyses where such parameters are stochastic using techniques from random graph theory. One can allow the $J_{i,j}$'s to depend on the characteristics of agent j as well as agent i, so that $J_{i,j} = J\left(\boldsymbol{Z}_i, \boldsymbol{Z}_j\right)$.

These assumptions are sufficient to characterize the distribution of aggregate choices as a function of the distribution of various microeconomic characteristics. As a preliminary, we make two algebraic manipulations. Observe first that we can, without loss of generality, replace the private deterministic utility function of each individual with a linear function,

$$u\left(\omega_i, \boldsymbol{Z}_i\right) = h_i \omega_i + k_i, \tag{6}$$

where $h_i = h\left(\boldsymbol{Z}_i\right)$ and $k_i = k\left(\boldsymbol{Z}_i\right)$ are chosen so that

$$h_i + k_i = u\left(1, \boldsymbol{Z}_i\right), \tag{7}$$

and

$$-h_i + k_i = u\left(-1, \boldsymbol{Z}_i\right). \tag{8}$$

This linearization is permissible since the new function coincides with the original utility function on the support of the individual choices. Hence, it does not readily generalize when more than two choices are available.

Second, we expand the social utility term (5), using $\omega_i^2 = \omega_j^2 = 1$, in that

$$S\left(\omega_i, \boldsymbol{Z}_i, \mu_i^e\left(\boldsymbol{\omega}_{-i}\right)\right) = \sum_{j \neq i} J_{i,j}\left(\omega_i E_i\left(\omega_j\right) - 1\right), \tag{9}$$

which makes clear the role of pairwise interactions between each individual choice and the expected choices of others. Notice that the $J_{i,j}$ is equal to the cross-partial derivative of the social utility function, in that

$$J_{i,j} = \frac{\partial^2 V\left(\omega_i, \boldsymbol{Z}_i, \mu_i^e\left(\boldsymbol{\omega}_{-i}\right)\right)}{\partial \omega_i \partial E_i\left(\omega_j\right)} = \frac{\partial^2 S\left(\omega_i, \boldsymbol{Z}_i, \mu_i^e\left(\boldsymbol{\omega}_{-i}\right)\right)}{\partial \omega_i \partial E_i\left(\omega_j\right)}, \tag{10}$$

which means that the function measures the strategic complementarity between individual choices and the expected choices of others. See Cooper and John (1988) for a general analysis of complementarities which provides many insights which will reappear in our framework. Unlike the standard formulation of complementarities, our interactions are driven by expectations of the behavior of others, rather than by their actual behavior.

The probability that individual i makes choice ω_i is equal to the probability that the utility of the choice exceeds that of $-\omega_i$,

$$\begin{aligned}&\mu\left(\omega_i \mid \boldsymbol{Z}_i, \mu_i^e\left(\boldsymbol{\omega}_{-i}\right)\right)\\ &\quad = \mu\left(V\left(\omega_i, \boldsymbol{Z}_i, \mu_i^e\left(\boldsymbol{\omega}_{-i}\right), \epsilon_i\left(\omega_i\right)\right) > V\left(-\omega_i, \boldsymbol{Z}_i, \mu_i^e\left(\boldsymbol{\omega}_{-i}\right), \epsilon_i\left(-\omega_i\right)\right)\right)\\ &\quad = \mu\Big(h_i\omega_i + \sum_{j \neq i} J_{i,j}\omega_i E_i\left(\omega_j\right) + \epsilon_i\left(\omega_i\right) > -\left(h_i\omega_i\right) - \sum_{j \neq i} J_{i,j}\omega_i E_i\left(\omega_j\right) + \epsilon_i\left(-\omega_i\right)\Big).\end{aligned} \tag{11}$$

Letting "~" denote "is proportional to," the logistic specification of the random utility terms means that this probability has the feature that

$$\mu\left(\omega_i \mid \boldsymbol{Z}_i, \mu_i^e\left(\boldsymbol{\omega}_{-i}\right)\right) \sim \exp\Big(\beta_i h_i \omega_i + \sum_{j \neq i} \beta_i J_{i,j}\omega_i E_i\left(\omega_j\right)\Big). \tag{12}$$

Since the random utility terms are independent across individuals, it must be the case that the joint set of choices obeys

$$\begin{aligned}&\mu\left(\boldsymbol{\omega} \mid \boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_I, \mu_1^e\left(\boldsymbol{\omega}_{-1}\right), \ldots, \mu_I^e\left(\boldsymbol{\omega}_{-I}\right)\right)\\ &\quad = \prod_i \mu\left(\omega_i \mid \boldsymbol{Z}_i, \mu_i^e\left(\boldsymbol{\omega}_{-i}\right)\right) \sim \prod_i \exp\Big(\beta_i h_i \omega_i + \sum_{j \neq i} \beta_i J_{i,j}\omega_i E_i\left(\omega_j\right)\Big).\end{aligned} \tag{13}$$

Equation (13) provides a general form for the joint probability measure of individual choices. It has the general form of a Gibbs measure, which is not coincidental. An important theorem in the statistical mechanics literature, due to Averintsev (1970) and Spitzer (1971), states that models of stochastic interactions of the type which have been outlined will generically possess probability measures with Gibbs representations.

To close the model, it is necessary to specify how expectations are determined. A natural special case of this model occurs when the agents all possess rational expectations, i.e.,

$$E_i\left(\omega_j\right) = E\left(\omega_j \mid \boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_I, E_k\left(\omega_l\right), \quad k = 1, \ldots, I,\ l = 1, \ldots, I\right). \tag{14}$$

The expectation operator on the right hand side is the mathematical expectation given by the equilibrium probability measure (13), when these same mathematical expectations are also the subjective expectations of each of the individual agents. This means that the expected values of each of the choices is constrained by a set of self-consistency conditions. In particular, the expected value of each of the individual choices for any set of beliefs will equal

$$E\left(\omega_i\right) = \tanh\left(\beta_i h_i + \sum_{j \neq i} \beta_i J_{i,j} E_i\left(\omega_j\right)\right), \tag{15}$$

and so rational expectations require that we replace the subjective expectations with their mathematical counterparts, i.e.,

$$E\left(\omega_i\right) = \tanh\left(\beta_i h_i + \sum_{j \neq i} \beta_i J_{i,j} E\left(\omega_j\right)\right). \tag{16}$$

These equations represent a continuous mapping of $[-1, 1]^I$ to $[-1, 1]^I$. Therefore, it is immediate from Brouwer's fixed point theorem that there is at least one fixed point solution, which implies Theorem 1.

Theorem 1. Existence of self-consistent equilibrium. *There exists at least one set of self-consistent expectations consistent with the binary choice model with interactions as specified by Equations (2), (4) and (5).*

By choosing particular specifications for the distribution of $\boldsymbol{Z}_i$ one can generate many of the models of binary choices with interactions which have appeared in the literature. Perhaps more important, these particular specifications illustrate the interesting aggregate properties of environments with interdependent decisionmaking.

We now consider some particular $J_{i,j}$ structures in order to develop more precise properties of the population's probabilistic behavior. Page (1997) provides a valuable analysis of the role of different interaction structures in generating different aggregate properties which supplements this discussion.

2.2. Global interactions

One version of the binary choice model assumes that interactions across individuals are global, in the sense that each individual assigns an identical weight to the expected choice of every other member of the population. Since a person always conforms to his own behavior, this is equivalent in terms of predicted behavior to assuming that an individual assigns a common weight to all persons including himself [Brock and Durlauf (1995)] and so we assume this for expositional purposes. Formally, given i,

$$J_{i,j} = \frac{J_i}{I} \ \forall\, j. \tag{17}$$

Notice that we normalize the global interaction term J_i by the population size I for analytical convenience. This specification seems especially plausible when individual groups are determined by large aggregates such as ethnicity, religion, or region.

Global interactions imply that an individual's choice is, outside of individual-specific characteristics, only influenced by his expectation of the average choice in the population, since Equation (17) implies that the social utility term may be rewritten (after taking the square) as

$$S\left(\omega_i, \boldsymbol{Z}_i, \mu_i^e\left(\boldsymbol{\omega}_{-i}\right)\right)=J_i\left(\omega_i E_i\left(\bar{\omega}_I\right)-1\right), \tag{18}$$

where $E_i\left(\bar{\omega}_I\right)$ denotes the subjective expectation of agent i of the population average $\bar{\omega}_I$. The joint probability measure for this case equals

$$\mu\left(\boldsymbol{\omega} \mid \boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_I, \mu_1^e\left(\boldsymbol{\omega}_{-1}\right), \ldots, \mu_I^e\left(\boldsymbol{\omega}_{-I}\right)\right) \sim \prod_i \exp\left(\beta_i h_i \omega_i+\beta_i J_i \omega_i E_i\left(\bar{\omega}_I\right)\right). \tag{19}$$

As for the general case, self-consistency requires that each individual's subjective belief concerning the average choice equals the mathematical expectation of the average choice,

$$E_i\left(\bar{\omega}_I\right)=E\left(\bar{\omega}_I \mid \boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_I, \mu_1^e\left(\boldsymbol{\omega}_{-1}\right), \ldots, \mu_I^e\left(\boldsymbol{\omega}_{-I}\right)\right) \forall\, i, \tag{20}$$

which combined with the expected value of each choice, Equation (15), means that for the global interactions model, any m is a self-consistent solution for the expected average choice level if it solves

$$m=\int \tanh\left(\beta(\boldsymbol{Z})\, h(\boldsymbol{Z})+\beta(\boldsymbol{Z})\, J(\boldsymbol{Z})\, m\right) \mathrm{d}F_{\boldsymbol{Z}}, \tag{21}$$

where $\mathrm{d}F_{\boldsymbol{Z}}$ denotes the empirical probability distribution of the observable individual characteristics. When each individual possesses identical observable characteristics h_i, β_i and J_i are constant across the population, which implies that this integral reduces to the equation

$$m=\tanh(\beta h+\beta J m). \tag{22}$$

This equation is easily analyzed and illustrates how multiple equilibria can emerge in interactions-based systems. Following the analysis in Brock and Durlauf (1995), these multiple equilibria can be described by Theorem 2.

Theorem 2. Number of equilibria in the binary choice model with interactions.

i. *If $\beta J>1$ and $h=0$, there exist three different values of m which solve Equation (22). One of these roots is positive, one root is zero, and one root is negative.*

ii. *If $\beta J>1$ and $h \neq 0$, there exists a threshold H (which depends on β and J) such that*

   a. *for $|\beta h|<H$, there exist three solutions m to Equation (22), one of which has the same sign as h, and the others possessing the opposite sign.*

b. for $|\beta h| > H$, there exists a unique solution m to Equation (22) with the same sign as h.

This theorem can be extended to the more general specification

$$m = \int \tanh\left(\beta h\left(\boldsymbol{Z}\right) + \beta Jm\right) \mathrm{d}F_{\boldsymbol{Z}}, \tag{23}$$

which differs from Equation (21) in that here β and J are assumed to be constant across individuals. The case of heterogeneous h_i's is of particular interest when considering the econometric implementation of the model.

In order to generalize our theorem, we define the function $R(\cdot)$ by

$$R\left(m\right) = \int \tanh\left(\beta h\left(\boldsymbol{Z}\right) + \beta Jm\right) \mathrm{d}F_{\boldsymbol{Z}}, \tag{24}$$

so that the integral can be treated as a function of m. Suppose that $\mathrm{d}F_{\boldsymbol{Z}}$ is symmetrically distributed with mean 0 and variance s and that $h\left(\boldsymbol{Z}\right)$ is symmetric about the origin. This implies that $R\left(0\right) = 0$ given $\tanh\left(-x\right) = -\tanh\left(x\right)$ and the assumed symmetry in h and $\mathrm{d}F_{\boldsymbol{Z}}$. Next, define

$$r\left(m\right) = \int \tanh'\left(\beta h\left(\boldsymbol{Z}\right) + \beta Jm\right) \mathrm{d}F_{\boldsymbol{Z}}. \tag{25}$$

Observe that

$$R'\left(0\right) = \beta J \int \tanh'\left(\beta h\left(\boldsymbol{Z}\right)\right) \mathrm{d}F_{\boldsymbol{Z}} = \beta J r\left(0\right) > 0. \tag{26}$$

This means that for sufficiently small βJ, $\beta J r\left(0\right) < 1$ but if $\beta J > r\left(0\right)^{-1}$then $R'\left(0\right) > 1$ and hence at least two new equilibria exist besides $m = 0$. On the other hand, note that for any pair m_1 and m_2

$$\left|R\left(m_1\right) - R\left(m_2\right)\right| \leqslant \beta J \left|m_1 - m_2\right|, \tag{27}$$

using Equation (23), the mean value theorem, and the facts that the tanh function is bounded between −1 and 1 and $\mathrm{d}F_{\boldsymbol{Z}}$ is a probability measure. If $\beta J < 1$, then this is a contraction mapping and there exists only one solution to Equation (23) in this case. Hence the $m = 0$ solution bifurcates into at least three solutions as βJ increases beyond 1. Notice that unlike the case of homogeneous h's, we have not ruled out the possibility that more than three equilibria exist. We summarize this as a corollary.

Corollary 1. Number of equilibria in binary choice model with global interactions and individual heterogeneity

If $h\left(\boldsymbol{Z}\right)$ is distributed symmetrically about the origin, then

i. If $\beta J < 1$, then the self-consistent equilibrium in Equation (23) is unique.

ii. If $\beta J > 1$, then there exist at least three self-consistent solutions to Equation (23).

2.3. Local interactions

Local interactions models typically assume that each agent interacts directly with only a finite number of others in the population. For each i, the set of j's with whom he has interactions is referred to as his neighborhood and is denoted by n_i. While residential neighborhoods have been a longstanding focus of the interactions literature, the models we analyze have much broader applicability.

In a local interactions model, the notion of neighborhood-level interactions is captured by a restriction on the interaction weights $J_{i,j}$ of the general form

$$J_{i,j} = 0 \text{ if } j \notin n_i. \tag{28}$$

Of course, the global interactions model can be treated as a special case of a neighborhoods model, one where all other members of the population are members of each i's neighborhood.

Depending on the application, the index i has been interpreted differently. For example, in Föllmer (1974) or Glaeser et al. (1996), $|i-j|$ measures the distance between individuals, whereas it is treated as an index of technological similarity as in Durlauf (1993). This allows one to construct a neighborhood for agent i by taking all agents within some fixed distance from i. (The distance can vary with direction.) This latter assumption is the source of the term "local." For purposes of analysis of finite systems, it is typical to locate actors on a torus so that distance can be defined symmetrically for all agents. (A 2-dimensional torus is formed out of a $k \times k$ lattice by connecting the east/west and north/south boundaries so as to ensure that each element of the resulting system has four nearest neighbors.) For agents located on a torus, one can rewrite social utility as

$$S\left(\omega_i, \boldsymbol{Z}_i, \mu_i^e\left(\boldsymbol{\omega}_{-i}\right)\right) = -E_i \sum_{j \in n_i} \frac{J_{i,j}}{2} \left(\omega_i - \omega_j\right)^2, \tag{29}$$

with associated individual probability measure

$$\mu\left(\omega_i \mid \boldsymbol{Z}_i, \mu_i^e\left(\boldsymbol{\omega}_{-i}\right)\right) \sim \exp\Big(\beta_i h_i \omega_i + \sum_{j \in n_i} \beta_i J_{i,j} \omega_i E_i\left(\omega_j\right)\Big), \tag{30}$$

and joint probability measure

$$\begin{aligned} &\mu\left(\boldsymbol{\omega} \mid \boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_I, \mu_1^e\left(\boldsymbol{\omega}_{-1}\right), \ldots, \mu_I^e\left(\boldsymbol{\omega}_{-I}\right)\right) \\ &\quad = \prod_i \mu\left(\omega_i \mid \boldsymbol{Z}_i, \mu_i^e\left(\boldsymbol{\omega}_{-i}\right)\right) \sim \prod_i \exp\Big(\beta_i h_i \omega_i + \sum_{j \in n_i} \beta_i J_{i,j} \omega_i E_i\left(\omega_j\right)\Big). \end{aligned} \tag{31}$$

A special case of the local interactions model occurs when local interactions are homogeneous, which means 1) all neighborhoods have the same size which we denote

N, and 2) within a neighborhood, all interaction weights are equal to a common J. In this special case, social utility will equal

$$S(\omega_i, \boldsymbol{Z}_i, \mu_i^e(\boldsymbol{\omega}_{-i})) = J\omega_i \sum_{j \in n_i} E_i(\omega_j), \tag{32}$$

where N denotes the number of members of a neighborhood. Under rational expectations it is immediate that one joint probability measure for agents' choices is

$$\mu(\boldsymbol{\omega}) \sim \exp\Big(\beta h \sum_i \omega_i + \beta N J \sum_i \omega_i E(\omega)\Big), \tag{33}$$

where

$$E(\omega) = \tanh(\beta h + \beta N J E(\omega)) = E(\omega_i) \ \forall\, i, \tag{34}$$

which implies the following theorem:

Theorem 3. Relationship between global and local interactions models. *Any equilibrium expected individual and average choice level m for the global interactions model is also an equilibrium expected individual and average choice in a homogeneous local interactions model.*

This result might initially appear odd, given the explicit local interaction structure of preferences. In fact, the equivalence is not surprising. When all expectations are identical, and the sample mean is required to equal the population mean, then agents are all implicitly connected to one another through the expectations formation process. To be clear, the local interactions model can exhibit equilibria which are different from that of the global case.

Focusing on the case where each individual is required to possess identical $E(\omega_i)$'s is not required by the logic of the local interactions model. There has been little work on the existence and characterization of asymmetric equilibrium $E(\omega_i)$'s, i.e., equilibria where the expected values differ across agents. Examples of asymmetric equilibria of this type may be found in Blume and Durlauf (1998b). A trivial example can be produced by taking two environments which exhibit global interactions and multiple equilibria and defining them as a common population.

Finally, it is worth noting that when interactions between decisions are all intertemporal, then the assumption of extreme-valued random utility increments can be dropped. The equilibrium properties of the dynamic models in this section can be recomputed under alternative probability densities such as probit which are popular in the discrete choice work. In fact, under the mean field analysis of global interactions, alternative specifications can incorporate probit or other densities as well. In both cases, the large scale properties of models under alternative error distributions are largely unknown.

2.4. Relationship to statistical mechanics

The models we have thus far outlined bear a close relationship to models in statistical mechanics. A standard question in statistical mechanics concerns how a magnet can exist in nature. A magnet is defined as a piece of iron in which a majority of the atoms are either spinning up or down. Since there is no physical reason why atoms should be more likely to spin up or down when considered in isolation, the existence of a natural magnet, which requires literally billions of atoms to be polarized towards one type of spin, would seem extraordinarily unlikely by the law of large numbers. As a result, statistical mechanics models are based on the primitive idea that the probability that one atom has a given spin is an increasing function of the number of atoms with the same spin within the atom's neighborhood. For the Ising model of ferromagnetism, the assumption is that atoms are arrayed on a 2- (or higher) dimensional integer lattice, so that

$$\begin{aligned}&\mu\left(\omega_i \mid \text{ spins of all other atoms in material}\right) = \\ &\mu\left(\omega_i \mid \ \omega_j \text{ such that } |i-j| = 1\right) \sim \exp\Big(\beta J \omega_i \sum_{|i-j|=1} \omega_j\Big).\end{aligned} \tag{35}$$

For the Curie–Weiss model, the physical interaction structure is assumed to be such that each atom's spin is probabilistically dependent on the average spin in the system, so that

$$\mu\left(\omega_i \mid \text{ spins of all other atoms in material}\right) \sim \exp\left(\beta J \omega_i \bar{\omega}\right). \tag{36}$$

Hence our models of binary choice with social interactions are mathematically quite similar to physical models of magnetism.

An important difference, however, does exist. While our socioeconomic model embeds pairwise interactions via the products of individual choices ω_i with the expected choices of others, the physical models are based upon conditional probabilities which depend on the products of the realized individual choices for all pairs of individuals. Interestingly, the physics literature has also dealt with expectations-based interactions. It turns out that models with interactions across realizations are extremely difficult to analyze, so physicists have developed what is referred to as a "mean-field approximation" to various ferromagnetism models. A mean-field approximation amounts to replacing certain terms in an original model with their mathematical expectation. Hence, the mean-field approximation for the conditional probability of the spin of a given atom for the Curie–Weiss model is

$$\mu\left(\omega_i\right) \sim \exp\left(\beta J \omega_i E\left(\bar{\omega}_I\right)\right), \tag{37}$$

which is of the same form as Equation (12) when agents possess identical $\boldsymbol{Z}_i$'s and $J_{i,j} = \frac{J}{I}$. Of course, what is an approximate model in a physical context is an exact model in the socioeconomic context we have been analyzing, at least given our behavior primitives. This difference occurs because our behavioral assumption is that individuals interact through their expectations of one another's behavior, rather than through realizations.

This last remark relates to a more general consideration in the use of statistical mechanics methods by social scientists. A basic conceptual difference exists between social and physical environments which contain interactions. Physical (and many mathematical) models of interactions typically take as primitives the conditional probabilities linking elements of a system, i.e., $\mu(\omega_1 \mid \boldsymbol{\omega}_{-1}), \ldots, \mu(\omega_I \mid \boldsymbol{\omega}_{-I})$. Analysis of the model considers the existence and (if so) properties of whatever joint probability measures are consistent with the conditional ones. In socioeconomic contexts, it is more natural to take preferences, beliefs, and technologies as primitives and from them determine what conditional probability relationships will hold. Hence, statistical mechanics and related models cannot be employed in socioeconomic contexts without determining what socioeconomic primitives will lead to a particular conditional probability representation. Further, the purposefulness of the objects of analysis in social science contexts also means that issues of the endogeneity of neighborhoods and the potential for the existence of institutions which coordinate collective action will naturally arise. These issues have no analog in physical contexts and are suggestive of the limitations in importing methods from physics into socioeconomic studies.

2.5. Social planning problem

Our analysis thus far has assumed that individual decisions are not coordinated. An alternative approach is to examine how decisions would be made when coordinated by a social planner. Beyond its use in developing welfare comparisons and developing contrasts with the noncooperative case, the social planner's solution may have empirical content in some contexts. As described by Coleman (1988, 1990; Chapter 12) the evolution of social capital, defined to include aspects of social structure which facilitate coordination across individuals and which may be embedded either in personal mores or organizations such as churches or schools, implies that in many types of social situations, coordinated behavior can emerge.

In order to do this, it is necessary to be more precise in the formulation of the underlying game played by members of the population. As before, we consider a population of I individuals each with payoff function $V\left(\omega_i, \boldsymbol{Z}_i, \mu_i^e(\boldsymbol{\omega}_{-i}), \epsilon_i(\omega_i)\right)$. The random functions $\epsilon_i(\cdot)$ are assumed to be observed by the members of the population, so that each agent i knows the realizations of $\epsilon_j(\cdot)\, \forall j \neq i$. We further assume that the distribution of these random components is described by Equation (2). Hence in terms of timing, nature draws the random functions $\epsilon_i(\cdot)$ and reveals them to the entire population. Second, players play the game G defined by

$$G = \left\{V\left(\omega_i, \boldsymbol{Z}_i, \mu_i^e(\boldsymbol{\omega}_{-i}), \epsilon_i(\omega_i)\right), \quad i = 1 \cdots I\right\}, \tag{38}$$

where $\mu_i^e(\boldsymbol{\omega}_{-i})$ denotes their beliefs about the behavior of other agents and is conditioned on nature's draw of the random functions.

With respect to this environment, an obvious benchmark is a perfect foresight Nash equilibrium. By this, we mean that each player knows the $\epsilon_i(\cdot)$ functions for every agent and forms beliefs about the resultant choices in the population $\mu_i^e(\boldsymbol{\omega}_{-i})$ which are confirmed in equilibrium. If each player is playing a pure strategy, this means that $\mu_i^e(\boldsymbol{\omega}_{-i}) = \boldsymbol{\omega}_{-i}$ so that a perfect foresight pure strategy equilibrium is a set of choices $\boldsymbol{\omega}$ such that for all i

$$\omega_i = \arg \max_{\gamma \in \{-1,1\}} V(\gamma, \boldsymbol{Z}_i, \boldsymbol{\omega}_{-i}, \epsilon_i(\gamma)). \tag{39}$$

For the analogous mixed strategy equilibrium, let $\boldsymbol{\pi}_i = \left(\pi_{i,-1}, \pi_{i,1}\right)$ denote the row vector of probability weights assigned by agent i to the two choices. Then $\boldsymbol{\Pi}_i = (\boldsymbol{\pi}_1 \cdots \boldsymbol{\pi}_I)$ denotes a perfect foresight Nash equilibrium if each $\boldsymbol{\pi}_i$ is consistent with

$$\begin{aligned}&\boldsymbol{\pi}_i = \arg\max_{\gamma_i} \gamma_{i,1} V(1, \boldsymbol{Z}_i, \boldsymbol{\Pi}_{-i}, \epsilon_i(1)) + \gamma_{i,-1} V(-1, \boldsymbol{Z}_i, \boldsymbol{\Pi}_{-i}, \epsilon_i(-1))\\ &\text{such that } \gamma_{i,-1}, \gamma_{i,1} \geqslant 0 \text{ and } \gamma_{i,-1} + \gamma_{i,1} = 1,\end{aligned} \tag{40}$$

so that agent i plays the mixture $\boldsymbol{\gamma}_i$ against the mixtures played by the other agents, $\boldsymbol{\Pi}_{-i} = (\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_{i-1}, \boldsymbol{\pi}_{i+1}, \ldots, \boldsymbol{\pi}_I)$. Mixture $\boldsymbol{\gamma}_i$ means that i chooses 1 with probability $\gamma_{i,1}$ and chooses -1 with probability $\gamma_{i,-1}$. It is a standard result that a mixed Nash equilibrium of this type will always exist, although a pure strategy Nash equilibrium may not.

Alternatively, a limited information Nash equilibrium can be characterized when agents make choices without knowledge of the $\epsilon_i(\cdot)$ functions for agents other than themselves. In terms of timing, one can think of agents forming beliefs $\mu_i^e(\boldsymbol{\omega}_{-i})$ before any $\epsilon_i(\cdot)$'s are realized, nature then drawing the $\epsilon_i(\cdot)$'s, revealing $\epsilon_i(\cdot)$ to agent i, and each i then choosing ω_i. For this case,

$$\omega_i = \arg \max_{\gamma \in \{-1,1\}} V(\gamma, \boldsymbol{Z}_i, \mu_i^e(\boldsymbol{\omega}_{-i}), \epsilon_i(\gamma)), \tag{41}$$

when $\mu_i^e(\boldsymbol{\omega}_{-i}) = \mu\left(\boldsymbol{\omega}_{-i} \mid \boldsymbol{Z}_j \forall j\right) \forall i$, so that each agents beliefs are consistent with the model. This is the equilibrium concept we have employed above.

In contrast to these noncooperative environments, we may characterize a social planner's perfect foresight problem as choosing $\boldsymbol{\omega}$ in order to maximize total utility in the population, i.e.,

$$\max_{\boldsymbol{\omega}} \sum_{i=1}^{I} V(\omega_i, \boldsymbol{Z}_i, \mu_i^e(\boldsymbol{\omega}_{-i}), \epsilon_i(\omega_i)). \tag{42}$$

From Equation (42), one can in principle compute quantities such as the expected average payoff under a social planner and contrast it with their counterparts under the two noncooperative environments.

In order to perform such a comparison, however, analytical tractability becomes a problem. To see this, notice that for our global interactions model, the social planner's problem becomes

$$\max_{\boldsymbol{\omega}} \sum_{i=1}^{I} \left(h_i \omega_i - \frac{J}{2} (\omega_i - \bar{\omega}_I)^2 + \epsilon_i (\omega_i) \right). \tag{43}$$

Unfortunately, $\sum_{i=1}^{I} \epsilon_i (\omega_i)$ is not independent and extreme value distributed over the 2^I possible configurations of $\boldsymbol{\omega}$, even though the individual $\epsilon_i (\omega_i)$'s are distributed that way. One way around this problem is, following Brock and Durlauf (1995), to replace this original social planner's problem with an approximate problem

$$\max_{\boldsymbol{\omega}} \sum_{i=1}^{I} \left(h_i \omega_i - \frac{J}{2} (\omega_i - \bar{\omega}_I)^2 + \epsilon_i^* (\omega_i) \right), \tag{44}$$

where $\epsilon (\boldsymbol{\omega}) \equiv \sum_{i=1}^{I} \epsilon_i^* (\omega_i)$ is itself extreme value distributed. One can require that the variance of the errors in the approximate social planner's problem equal those in the original problem in order to achieve some calibration between the two problems.

Under our assumption on $\epsilon (\boldsymbol{\omega})$, one may use Equation (44) to show that the probability measure characterizing the joint choice of $\boldsymbol{\omega}$ equals

$$\mu (\boldsymbol{\omega}) = \frac{\exp \left(\beta \left(\sum_{i=1}^{I} h_i \omega_i + \frac{J}{2I} \left(\sum_{i=1}^{I} \omega_i \right)^2 \right) \right)}{\sum_{\nu_1 \in \{-1,1\}} \cdots \sum_{\nu_I \in \{-1,1\}} \exp \left(\beta \left(\sum_{i=1}^{I} h_i \nu_i + \frac{J}{2I} \left(\sum_{i=1}^{I} \nu_i \right)^2 \right) \right)}. \tag{45}$$

In order to analyze this probability measure, which is known in the statistical mechanics literature as the Curie–Weiss model, it is necessary to eliminate the $\left(\sum_{i=1}^{I} \omega_i \right)^2$ terms in Equation (45). This calculation is complicated and may be found in the Appendix; further analysis appears in Brock (1993). A result currently exists only for the case $h_i = h$ and only for the large economy limit. However, Amaro de Matos and Perez (1991) suggest that for the large economy limit generalization to heterogeneous h_is is possible. The Appendix verifies Theorem 4.

Theorem 4. Expected average choice under social planner for binary choice model with interactions. *Let m^* denote the root of $m^* = \tanh(\beta h + \beta J m^*)$ with the same sign as h. If Equation (39) characterizes the joint distribution of individual choices as determined by a social planner, then*

$$\lim_{I \Rightarrow \infty} E (\bar{\omega}_I) = m^*. \tag{46}$$

One aspect of this theorem is intuitive, in that a planner would choose that average choice level in which the interaction effects and the private deterministic utility comparisons work together. What is perhaps surprising is that the social planner's equilibrium is sustainable as an equilibrium in the limited information noncooperative environment. However, this result is somewhat special to the functional form originally assumed for individual deterministic social utility. If the original social utility term had been $J\omega_i E(\bar{\omega}_I)$, then the noncooperative equilibrium average choice level would be the same as for the case we have studied, but the analogous social planner's problem would choose that root of $m^* = \tanh(\beta h + 2\beta J m^*)$ with the same sign as h [Brock and Durlauf (1995)], which would mean it is not supportable in the limited information noncooperative environment.

2.6. Linear-in-means model

Much of the empirical work on interaction effects has assumed that the behavior variable ω_i has continuous support and depends linearly on various individual and neighborhood effects. These assumptions permit a researcher to use ordinary least squares methods, which will be discussed below. While these empirical papers generally do not consider what decision problems generate their econometric specifications, it is straightforward to do so. For a trivial example, suppose that an individual solves

$$\max_{\omega_i \in (-\infty, \infty)} -\tfrac{1}{2}\left(\omega_i - \omega_i^*\right)^2, \tag{47}$$

where ω_i^* is a reference behavior level to which individual i prefers to conform. When this reference behavior level equals $h_i + J E_i(\bar{\omega}_I) + \epsilon_i$, it is immediate that

$$\omega_i = h_i + J E_i(\bar{\omega}_I) + \epsilon_i. \tag{48}$$

This is the type of equation studied by Manski (1993a,b), Moffitt (1998), Duncan and Raudenbusch (1998), among others.

3. Identification: basic issues

In this section, we describe the identification of interactions-based models in cross-sections. Identification is a concern in these cases because of the likelihood that group versus individual determinants of individual behavior are correlated. Hauser (1970) provides an early and clever analysis of how these correlations can, if not properly accounted for, lead to spurious inferences. We recommend this paper as an example of how powerful intuitive reasoning (as well as good common sense) can complement and foreshadow formal analysis.

Manski (1993a,b, 1997) has pioneered the study of the identification of interaction effects, and we will follow his treatment closely. In his work, Manski distinguishes between three explanations for correlated behavior within groups:

> "*endogenous effects*, wherein the propensity of an individual to behave in some way varies with the behaviour of the group... *exogenous (contextual)* effects, wherein the propensity of an individual to behave in some way varies with the exogenous characteristics of a group... *correlated* effects, wherein individuals in the same group tend to behave similarly because they have similar individual characteristics or face similar institutional environments" [Manski (1993a, p. 532)]

The treatment of identification problems in terms of the ability to distinguish these different effects in data seems to us very useful and so we employ it throughout.

For purposes of discussing identification and other econometric aspects of interactions-based models, we begin with a baseline set of data assumptions which will apply both to the binary choice model and to the linear-in-means model. We assume that the econometrician has available a set of observations on I individuals. We assume that each individual is drawn randomly from a set of neighborhoods. Within each neighborhood, all interactions are global. For notational purposes, we denote individuals as i and the neighborhood (which means the set of other individuals who influence i through interactions) as $n(i)$. We assume that our original vector $\boldsymbol{Z}_i$ can be partitioned into an r-length vector of individual-specific observables $\boldsymbol{X}_i$ and an s-length vector of exogenously determined neighborhood observables $\boldsymbol{Y}_{n(i)}$ associated with each individual in the sample. This will allow us to replace the private utility component h_i in our theoretical discussion with a linear specification

$$h_i = k + c'\boldsymbol{X}_i + d'\boldsymbol{Y}_{n(i)}. \tag{49}$$

Notice that this specification means that none of the individual-specific observables $\boldsymbol{X}_i$ or neighborhood observables $\boldsymbol{Y}_{n(i)}$ contains a constant term. We will maintain this assumption throughout. Within a neighborhood, all interactions are assumed to be global and symmetric, so that there is a single parameter J which indexes interactions.

Recall that $m^e_{n(i)}$ is agent i's subjective expectation of the average choice in neighborhood $n(i)$. In the subsequent discussion, it will be useful to distinguish between $m^e_{n(i)}$ and $m_{n(i)}$, the mathematical expectation of the average choice in a neighborhood under self-consistency. (We will specify the information sets under which self-consistency is calculated below.) The reason for this is that we will have need to distinguish between the data in a statistical exercise and the mathematical solution to a model. Of course, $m^e_{n(i)} = m_{n(i)}$ is part of our maintained assumption in the analysis, so there is no loss of generality in doing this. For purposes of discussion of identification, we are therefore either implicitly assuming that the neighborhoods are arbitrarily large, so that the neighborhood sample average can be used in place of the expected value or that accurate survey data are available. We finally assume that the errors are independent across individuals and that

$\mu(\epsilon_i(\omega_i) - \epsilon_i(-\omega_i) \mid \boldsymbol{X}_i, \boldsymbol{Y}_{n(i)}, m^e_{n(i)}) = \mu(\epsilon_i(\omega_i) - \epsilon_i(-\omega_i))$ for the binary choice model, and $E(\epsilon_i \mid \boldsymbol{X}_i, \boldsymbol{Y}_{n(i)}, m^e_{n(i)}) = 0$ for the linear-in-means model.

Our strategy of using individual level data has an important advantage: to the extent that the parameters of the individual model are identified, one can infer whether or not multiple equilibria exist with respect to population aggregates. This can be done without consideration of an equilibrium selection rule because population aggregates are always treated as independent variables in the analysis. Hence we can circumvent some of the problems described in Jovanovic (1989).

3.1. Binary choice

For the binary choice model, we consider the identification based on a naive estimator of the parameters of the model. By naive, we refer to the case where a logistic regression is computed which does not impose the relationships between neighborhood means. In this case, the conditional likelihood function for the set of individual choices will have a standard logistic form. Using our theoretical model of global interactions (and exploiting symmetry of the logistic density function), the likelihood is

$$\begin{aligned} &L\left(\boldsymbol{\omega}_I \mid \boldsymbol{X}_i, \boldsymbol{Y}_{n(i)}, m^e_{n(i)} \forall\, i\right) \\ &\quad= \prod_i \mu\left(\omega_i = 1 \mid \boldsymbol{X}_i, \boldsymbol{Y}_{n(i)}, m^e_{n(i)}\right)^{\frac{1+\omega_i}{2}} \cdot \mu\left(\omega_i = -1 \mid \boldsymbol{X}_i, \boldsymbol{Y}_{n(i)}, m^e_{n(i)}\right)^{\frac{1-\omega_i}{2}} \\ &\quad\sim \prod_i \Big(\exp\left(\beta k + \beta c' \boldsymbol{X}_i + \beta d' \boldsymbol{Y}_{n(i)} + \beta J m^e_{n(i)}\right)^{\frac{1+\omega_i}{2}} \\ &\qquad\qquad \cdot \exp\left(-\beta k - \beta c' \boldsymbol{X}_i - \beta d' \boldsymbol{Y}_{n(i)} - \beta J m^e_{n(i)}\right)^{\frac{1-\omega_i}{2}} \Big). \end{aligned} \tag{50}$$

As is standard for logistic models, the complete set of model parameters is not identified as k, c', d' and J are each multiplied by β. We therefore proceed under the normalization $\beta = 1$.

The reason that identification is a concern in a model like this is the presence of the term $m^e_{n(i)}$ in the likelihood function. Since this term embodies a rationality condition, it is a function of other variables in the likelihood function. Specifically, we assume that

$$m^e_{n(i)} = m_{n(i)} = \int \tanh\left(k + c'\boldsymbol{X} + d'\boldsymbol{Y}_{n(i)} + J m_{n(i)}\right) \mathrm{d}F_{\boldsymbol{X} \mid \boldsymbol{Y}_{n(i)}}. \tag{51}$$

Here $F_{\boldsymbol{X} \mid \boldsymbol{Y}_{n(i)}}$ denotes the conditional distribution of $\boldsymbol{X}$ in neighborhood $n(i)$ given the neighborhood characteristics $\boldsymbol{Y}_{n(i)}$. What this means is that each agent is assumed to form the conditional probabilities of the individual characteristics in a neighborhood given the aggregates which determine his or her payoffs. Since one can always add elements of $\boldsymbol{Y}_{n(i)}$ with zero coefficients to the payoff equation for agents, this is without loss of generality.

Rather than prove identification for the particular case where the theoretical model is logistic [see McFadden (1974) and Amemiya (1985; Chapter 9) for proofs for this case] we prove identification for an arbitrary known distribution function for the random payoff terms. Specifically, we assume that the conditional probability of individual i's choice can be written as

$$\mu\left(\epsilon(\omega_i) - \epsilon(-\omega_i) \leqslant z \mid \boldsymbol{X}_i, \boldsymbol{Y}_{n(i)}, m^e_{n(i)}\right) = F\left(z \mid k + c'\boldsymbol{X}_i + d'\boldsymbol{Y}_{n(i)} + Jm^e_{n(i)}\right), \tag{52}$$

where F is a known probability distribution function that is continuous and strictly increasing in z.

We consider identification based on a naive estimator of the parameters of the model. By naive, we refer to the situation where parameter estimates for the model are computed which do not impose the rational expectations condition between neighborhood means and neighborhood characteristics, but rather uses these variables as regressors. Hence, we assume that $m^e_{n(i)}$ is known to the researcher; see discussion below for the case when $m^e_{n(i)}$ is not observable.

To formally characterize identification, we employ the following notation. Define $\text{supp}(\boldsymbol{X}, \boldsymbol{Y}, m^e)$ as the joint support of the distribution of $(\boldsymbol{X}_i, \boldsymbol{Y}_{n(i)}, m^e_{n(i)})$. Intuitively, the definition of identification we employ says that a model is identified if there do not exist two distinct sets of parameter values each of which produces (for all subsets of $\boldsymbol{X}$ and $\boldsymbol{Y}$ which occur with positive probability) identical probabilities for individual choices and which are also self-consistent.

Definition. *Global identification in the binary choice model with interactions and self-consistent expectations*: The binary choice model is globally identified if for all parameter pairs (k, c, d, J) and $(\bar{k}, \bar{c}, \bar{d}, \bar{J})$

$$k + c'\boldsymbol{X}_i + d'\boldsymbol{Y}_{n(i)} + Jm^e_{n(i)} = \bar{k} + \bar{c}'\boldsymbol{X}_i + \bar{d}'\boldsymbol{Y}_{n(i)} + \bar{J}m^e_i, \tag{53}$$

and

$$\begin{aligned} m^e_{n(i)} &= m_{n(i)} \\ &= \int \omega_i \mathrm{d}F\left(\omega_i \mid k + c'\boldsymbol{X} + d'\boldsymbol{Y}_{n(i)} + Jm_{n(i)}\right) \mathrm{d}F_{\boldsymbol{X} \mid \boldsymbol{Y}_{n(i)}} \\ &= \int \omega_i \mathrm{d}F\left(\omega_i \mid \bar{k} + \bar{c}'\boldsymbol{X} + \bar{d}'\boldsymbol{Y}_{n(i)} + \bar{J}m_{n(i)}\right) \mathrm{d}F_{\boldsymbol{X} \mid \boldsymbol{Y}_{n(i)}} \\ &\forall \left(\boldsymbol{X}_i, \boldsymbol{Y}_{n(i)}, m^e_{n(i)}\right) \in \text{supp}\left(\boldsymbol{X}, \boldsymbol{Y}, m^e\right), \end{aligned} \tag{54}$$

imply that $(k, c, d, J) = (\bar{k}, \bar{c}, \bar{d}, \bar{J})$.

In order to establish conditions under which identification can hold we follow the argument in Manski (1988), Proposition 5, and state the following Proposition, whose proof appears in the Appendix. The assumptions we make are clearly sufficient rather than necessary; weakening the assumptions is left to future work. In interpreting the assumptions, note that Assumption *i* is the one used by Manski to identify this model when there are no endogenous effects, i.e., if J is known a priori to be 0. The assumption, of course, does nothing more than ensure that the individual and contextual regressors are not linearly dependent. The additional assumptions are employed to account for the fact that $m_{n(i)}$ is a nonlinear function of the contextual effects.

Theorem 5. Sufficient conditions for identification to hold in the binary choice model with interactions and self-consistent beliefs. *Assume*

i. $supp(\boldsymbol{X}_i, \boldsymbol{Y}_{n(i)})$ is not contained in a proper linear subspace of R^{r+s}.
ii. $supp(\boldsymbol{Y}_{n(i)})$ is not contained in a proper linear subspace of R^s.
iii. No element of $\boldsymbol{X}_i$ or $\boldsymbol{Y}_{n(i)}$ is constant.
iv. There exists at least one neighborhood n_0 such that conditional on $\boldsymbol{Y}_{n_0}$, $\boldsymbol{X}_i$ is not contained in a proper linear subspace of R^r.
v. None of the regressors in $\boldsymbol{Y}_{n(i)}$ possesses bounded support.
vi. $m_{n(i)}$ is not constant across all neighborhoods n.
Then, (k, c, d, J) is identified relative to any distinct alternative $(\bar{k}, \bar{c}, \bar{d}, \bar{J})$.

3.2. Linear-in-means model

Identification in the binary choice model with interactions can be contrasted with the case of the analogous linear-in-means model,

$$\omega_i = k + c'\boldsymbol{X}_i + d'\boldsymbol{Y}_{n(i)} + Jm^e_{n(i)} + \epsilon_i. \tag{55}$$

The unique self-consistent solution $m_{n(i)}$ for the linear-in-means model is easily seen, by applying an expectations operator to both sides of the individual behavioral equation, to be

$$m_{n(i)} = \frac{k + c'E\left(\boldsymbol{X}_i \mid \boldsymbol{Y}_{n(i)}\right) + d'\boldsymbol{Y}_{n(i)}}{1-J}, \tag{56}$$

where $E\left(\boldsymbol{X}_i \mid \boldsymbol{Y}_{n(i)}\right)$ denotes the expected value of the individual controls given the neighborhood characteristics. Hence, following the argument in Manski (1993a,b), one can construct a reduced form expression for individual choices,

$$\omega_i = \frac{k}{1-J} + c'\boldsymbol{X}_i + \frac{J}{1-J}d'\boldsymbol{Y}_{n(i)} + \frac{J}{1-J}c'E\left(\boldsymbol{X}_i \mid \boldsymbol{Y}_{n(i)}\right) + \epsilon_i. \tag{57}$$

In this equation, we have $2r+s+1$ regressors and $r+s+2$ parameters. The possibility for identification in this model therefore will depend on which, if any, of the regressors in the reduced form are linearly independent (i.e., their variance covariance matrix is of full rank). For example if $E\left(\boldsymbol{X}_i \mid \boldsymbol{Y}_{n(i)}\right)$ is linearly dependent on $\boldsymbol{Y}_{n(i)}$, then it is obvious that the model parameters are not identified. More generally, it is necessary for identification that the dimension of the linear space spanned by the regressors is at least equal to the number of structural parameters, i.e., $r+s+2$; otherwise, one cannot map the reduced form coefficients back to the structural parameters. Hence, one can state the following theorem.

Theorem 6. Necessary conditions for identification in the linear-in-means model with interactions and self-consistent beliefs. *In the linear-in-means model it is necessary for identification of the model's parameters that*

i. The dimension of the linear space spanned by elements of $\left(1, \boldsymbol{X}_i, \boldsymbol{Y}_{n(i)}\right)$ *is* $r+s+1$.
ii. The dimension of the linear space spanned by the elements of $(1, \boldsymbol{X}_i, \boldsymbol{Y}_{n(i)}, E$ $(\boldsymbol{X}_i \mid \boldsymbol{Y}_{n(i)}))$ *is at least* $r+s+2$.

Notice that the conditions of this theorem, while analogous to those in the theorem for identification in the binary choice model, are now necessary and not sufficient. This is because sufficient conditions will depend on the model parameters. For example, if $c = \boldsymbol{0}$, then the fact that $E\left(\boldsymbol{X}_i \mid \boldsymbol{Y}_{n(i)}\right)$ is linearly independent of the regressors $\boldsymbol{X}_i$ and $\boldsymbol{Y}_{n(i)}$ will not eliminate collinearity of $m_{n(i)}$ and $\boldsymbol{Y}_{n(i)}$ in the structural equation (55) and hence will leave only $s+1$ regression coefficients in the reduced form available to identify k, J and d, which is not enough.

This theorem is an extension of Manski's (1993a,b) result on the nonidentifiability of contextual versus endogenous effects. Manski's analysis assumes that there is a one-to-one correspondence between the individual control variables $\boldsymbol{X}_i$ and the neighborhood control variables $\boldsymbol{Y}_{n(i)}$ so that for any individual-level variable that influences behavior, the neighborhood average of that variable also influences behavior. For example, if one controls for individual education, one also controls for average neighborhood education. In this case, $E\left(\boldsymbol{X}_i \mid \boldsymbol{Y}_{n(i)}\right) = \boldsymbol{Y}_{n(i)}$. Hence, $m_{n(i)}$ is linearly dependent on $\boldsymbol{Y}_{n(i)}$ and so the model is not identified. Notice as well that the Theorem requires that $E\left(\boldsymbol{X}_i \mid \boldsymbol{Y}_{n(i)}\right)$ is a nonlinear function of $\boldsymbol{Y}_{n(i)}$; this is analogous to the condition for identification of some interaction effect in Manski (1993a,b), Proposition 1 and Corollary. (Manski's results have to do with the identification of either an endogenous or contextual effect in the presence of individual effects, but does not allow for identification between these two effects, whereas our result gives conditions under which the two group effects can be distinguished.)

Why is there this difference between the binary choice and the linear-in-means frameworks? The answer is that the binary choice framework imposes a nonlinear relationship between the group characteristics and the group behaviors whereas the linear-in-means model (of course) does the opposite. Intuitively, suppose that one moves an individual from one neighborhood to another and observes the differences in his behavior. If the characteristics and behaviors of the neighborhoods always move in proportion as one moves across neighborhoods, then clearly one could not determine the respective roles of the characteristics as opposed to the behavior of the group in determining individual outcomes. This can never happen in the logistic binary choice case given that the expected average choice must be bounded between -1 and 1. So, for example, as one moves across a sequence of arbitrarily richer communities, the percentage of high school graduates cannot always increase proportionately with income.

One can develop analogous identification conditions for alternative information assumptions in the linear-in-means model. For example, suppose that $\bar{\boldsymbol{X}}_{n(i)}$, the sample average of the individual characteristics in neighborhood $n(i)$, is known to all members of the neighborhood. In this case,

$$m_{n(i)} = \frac{k + c'\bar{\boldsymbol{X}}_{n(i)} + d'\boldsymbol{Y}_{n(i)}}{1 - J}. \tag{58}$$

This equation makes clear that if the elements of $\bar{\boldsymbol{X}}_{n(i)}$ lie in the linear space spanned by $\boldsymbol{Y}_{n(i)}$, then the linear-in-means model will not be identified. Hence, we have the following corollary.

Corollary 2. Necessary conditions for identification in the linear-in-means model when $\boldsymbol{Y}_{n(i)}$ and $\bar{\boldsymbol{X}}_{n(i)}$ are observable

If $\boldsymbol{Y}_{n(i)}$ and $\bar{\boldsymbol{X}}_{n(i)}$ are observable, then a necessary condition for identification in the linear-in-means model is that the dimension of the linear space spanned by $(1, \boldsymbol{X}_i, \boldsymbol{Y}_{n(i)}, \bar{\boldsymbol{X}}_{n(i)})$ is at least $r + s + 2$.

Operationally, this corollary means that for the full information case, one needs one individual variable whose neighborhood level average is not an element of the individual behavioral equation. This average can then be used to instrument $m^e_{n(i)}$.

3.3. Instruments for unobservable expectations

The identification condition for the linear-in-means model suggests a set of instruments which may be used when $m^e_{n(i)}$ is not observable, is measured with error, etc. Specifically, replacing $m^e_{n(i)}$ with the projection of $\bar{\omega}_{n(i)}$, the sample average of behaviors in neighborhood $n(i)$ onto $H\left(\boldsymbol{Y}_{n(i)}, E\left(\boldsymbol{X}_i \mid \boldsymbol{Y}_{n(i)}\right)\right)$, where $H(a, b)$ denotes the Hilbert space generated by the elements of vectors a and b, will not affect our identification results so long as $\dim(H(\boldsymbol{Y}_{n(i)}, E(\boldsymbol{X}_i \mid \boldsymbol{Y}_{n(i)})) \ominus H(\boldsymbol{Y}_{n(i)})) > 0$, where for Hilbert spaces I and G such that $G \subseteq I$, $I \ominus G$ denotes the Hilbert space generated by those elements of I that are orthogonal to all elements of G. An analogous procedure will apply when $\bar{\boldsymbol{X}}_{n(i)}$ is observable to individuals.

Of course, this assumes that the researcher has prior knowledge of what individual-level variables affect behavior when their neighborhood averages do not; otherwise, it would be the case that $H\left(E\left(\boldsymbol{X}_i \mid \boldsymbol{Y}_{n(i)}\right)\right) \subseteq H\left(\boldsymbol{Y}_{n(i)}\right)$ and so may be susceptible to Sims' (1980) classic critique of "incredible" identifying restrictions; see Freedman (1991) for a similar critique of the sorts of regressions we describe here. The point remains, however, that identification in the linear-in-means model depends on the same classical conditions as does identification in general simultaneous equations models, as initially recognized by Moffitt (1998).

At the same time, we would argue that the issue of omitted variables is far from insuperable. Both the social psychology and sociology literatures have focused a great deal of attention as to which types of individual and group control variables are most appropriate for inclusion in individual level regressions through the determination of which variables seem to be proximate versus ultimate causes of individual behavior. Indeed, it is this distinction which is the basis of path analysis [Blau and Duncan (1967)]; see Sampson and Laub (1995) for what we consider a persuasive example of such a study. In general, we find it likely that these literatures will be able to identify examples of individual variables whose group average analogs are not proximate causes of behavior, and hence are available as instruments. While these literatures are often not driven by formal statistical modelling and further subjected to Sims/Freedman-type critiques [e.g., Freedman (1991)] when formal techniques are employed, this hardly means that these literatures are incapable of providing useful insights. In this respect, we find arguments to the effect that because an empirical relationship has been established without justification for auxiliary assumptions such as linearity, exogeneity of certain variables, etc., one can ignore it, to be far overstated. In our view, empirical work establishes greater or lesser degrees of plausibility for different claims about the world and therefore the value of any study should not be reduced to a dichotomy between full acceptance or total rejection of its conclusions. Hence the determination of the plausibility of any exclusion restriction is a matter of degree and dependent on its specific context, including the extent to which it has been studied.

3.4. Identification of individual versus neighborhood contextual effects

We now consider in more detail what is involved for identification of some type of neighborhood effects. What we mean is the following. Suppose that one wishes to determine whether any type of neighborhood effect exists, without distinguishing between endogenous and contextual effects, hence the only regressors in the model are a constant, $\boldsymbol{X}_i$, and $\boldsymbol{Y}_{n(i)}$. Operationally, we define this as determining whether, for a statistical model which only includes contextual effects as controls, the parameters on these contextual effects are identified[3]. A Corollary of the general identification Theorems 5 and 6 highlights the two conditions necessary to distinguish individual versus neighborhood contextual effects. A related result may be found in Manski (1993a,b, corollary, p. 535).

Corollary 3. Identification of individual versus neighborhood effects in the binary choice model and linear-in-means model with global interactions

In either the binary choice or the linear-in-means models with global interactions, a necessary condition for the identification of some neighborhood effect is that the dimension of the linear space spanned by the elements of $\left(1, \boldsymbol{X}_i, \boldsymbol{Y}_{n(i)}\right)$ is of higher dimension than the linear space spanned by the elements of $(1, \boldsymbol{X}_i)$.

[3] When $\boldsymbol{Y}_{n(i)} = E\left(\boldsymbol{X}_i \mid i \in n(i)\right)$ the corollary can be interpreted as applying to identification of a group effect for the reduced form of the linear-in-means model.

While the corollary is trivial, in that it is nothing more than the statement that in order to identify some sort of neighborhood effect some combination of the expected neighborhood effects must be linearly independent of the individual controls, it does have some economic content. Suppose that individuals are sorted into neighborhoods on the basis of an individual characteristic and that the neighborhood average of this same characteristic is what constitutes the relevant contextual effect. What this means is that there exists a neighborhood assignment rule $\xi(\cdot)$ which relates individual characteristics to neighborhood characteristics such that nonidentification requires that (assuming that the set of neighborhood characteristics and the set of individual characteristics are each internally linearly independent) there exists some linear combination of individual characteristics which is equal to some linear combination of neighborhood characteristics, i.e., there exist weights α and γ such that

$$\kappa + \alpha' \boldsymbol{X}_i = \gamma' \boldsymbol{Y}_{n(i)}. \tag{59}$$

But if this is so, then if individual observations are chosen randomly from the neighborhood, it must be the case that individuals are perfectly segregated across neighborhoods with respect to the composite individual characteristic $\kappa + \alpha' \boldsymbol{X}_i$. This is an extremely strong condition on the neighborhood sorting rule, ruling out any noise in the sorting process, and is in our judgment implausible. Hence our interpretation of the identification corollary is that empirical researchers should feel confident that individual versus neighborhood effects can be at least in principle distinguished. To be clear however, this does not mean that data sets drawn from highly segregated communities are not a problem; rather the same reasoning we have applied suggests how segregation can lead to large standard errors for the estimated parameters of the model.

3.5. Nonlinear-in-means model

The differences between the binary choice and linear-in-means models suggest that nonlinearity has a fundamental effect on the identification problem. McManus (1992) provides a number of general results which indicate that lack of identification is a nongeneric phenomenon in nonlinear contexts; these contexts do not include self-consistency conditions of the type which created the identification problem in the linear-in-means model. (A property is generic to a topological space of objects if it holds for an open, dense subset of the space.) McManus' analysis relies on some results from differential topology, which are beyond the scope of this chapter.

Nevertheless, it is possible to demonstrate a basic role for nonlinearity in identifying the parameters of interactions-based models by examining deviations from the linear-in-means model we have studied. Suppose that the individual behavioral equation is

$$\omega_i = k + c' \boldsymbol{X}_i + d' \bar{\boldsymbol{X}}_{n(i)} + J m^e_{n(i)} + \epsilon_i. \tag{60}$$

Here, the contextual effects $\bar{\boldsymbol{X}}_{n(i)}$ are averages of the individual controls $\boldsymbol{X}_i$, so we know that this model is not identified by Theorem 6. In the spirit of McManus (1992), we wish to make precise the idea that for the class of models, when the model is not linear in $m^e_{n(i)}$ but rather is linear in a function of $m^e_{n(i)}$, lack of identification is pathological.

To do this, let $g(m)$ be a C^2 function such that g is nonlinear in m and let

$$G\left(m^e_{n(i)}\right) = m^e_{n(i)} + \xi g\left(m^e_{n(i)}\right), \tag{61}$$

represent a class of functions which are perturbations around the linear function $m^e_{n(i)}$. We consider the nonlinear-in-means model

$$\omega_i = k + c'\boldsymbol{X}_i + d'\bar{\boldsymbol{X}}_{n(i)} + J\,G\left(m^e_{n(i)}\right) + \epsilon_i. \tag{62}$$

Associated with this equation is a conditional mean function H

$$H\left(\boldsymbol{X}_i, \bar{\boldsymbol{X}}_{n(i)}, m^e_{n(i)}\right) = k + c'\boldsymbol{X}_i + d'\bar{\boldsymbol{X}}_{n(i)} + J\,G\left(m^e_{n(i)}\right). \tag{63}$$

For this model, self-consistency of $m^e_{n(i)}$ requires

$$m^e_{n(i)} = m_{n(i)} = k + \left(c' + d'\right)\bar{\boldsymbol{X}}_{n(i)} + J\,G\left(m_{n(i)}\right). \tag{64}$$

Our goal is to determine whether the model with $\xi = 0$ is special in terms of nonidentifiability of the parameters in Equation (60). In doing so, we will assume that when there are multiple solutions to this equation, there is a selection rule which selects a particular solution $m_{n(i)}$ so that the observed $m_{n(i)} = m\left(\bar{\boldsymbol{X}}_{n(i)}\right)$.

In analyzing this equation, we will work with a notion of local identification. The model Equations (61–64) define a "structure" for each particular parameter vector $A = (k, c, d, J)$. We focus here on identification at the level of the conditional mean function (63). Following Rothenberg (1971) or McManus (1992), we say a parameter point A_0 is locally identified if it fulfills the following definition. In our context, this condition is equivalent to requiring that the gradient vector of Equation (63) with respect to A to have full rank.

Definition. *Local identification in the nonlinear-in-means model with interactions and self-consistent beliefs*: For the model described by Equations (61–64), the parameter vector A_0 is locally identified if there exists an open neighborhood N_{A_0} of A_0 such that no other parameter vector in N_{A_0} gives the same conditional mean in Equation (63) and such that the self-consistency condition Equation (64) holds as well.

The concept of local identifiability has value as argued in Rothenberg (1971, p. 578), as

> "It is natural to consider the concept of local identification. This occurs when there may be a number of observationally equivalent structures but they are isolated from each another."

Rothenberg (1971) demonstrates that there is a close connection between local identification and the full rank assumption of particular derivative matrices of a likelihood function. In our context, this means that one must show that the gradient of the conditional mean function with respect to A is of full rank. In addition, we need to account for the self-consistency condition in the sense that the full rank condition must hold when the gradient is evaluated at a solution $m_{n(i)}$ to the self-consistency condition. The following Theorem is verified in the Appendix.

Theorem 7. Local identifiability for models in a neighborhood of the linear-in-means model. *Assume*

i. *$supp\left(\bar{\boldsymbol{X}}_{n(i)}\right)$ is not contained in a proper linear subspace of R^r.*

ii. *There exists at least one neighborhood n_0 such that conditional on $\bar{\boldsymbol{X}}_{n_0}$, $\boldsymbol{X}_i$ is not contained in a proper linear subspace of R^r.*

iii. $J \neq 1$.

iv. *The population data $\left\{\bar{\boldsymbol{X}}_{n(i)}, \boldsymbol{X}_i, m_{n(i)}\right\}$ is such that there is an open set O such that $m\left(\bar{\boldsymbol{X}}_{n(i)}\right)$ is differentiable on O and nonconstant on O. Further, there are two distinct values in O, call them $\bar{\boldsymbol{X}}_1$ and $\bar{\boldsymbol{X}}_2$, such that $m_1 = m\left(\bar{\boldsymbol{X}}_1\right) \neq m_2 = m\left(\bar{\boldsymbol{X}}_2\right)$ and $\frac{dg(m_1)}{dm} \neq \frac{dg(m_2)}{dm}$.*

Then there exists an open neighborhood N of $\xi = 0$, such that $\forall\ \xi \in N - \{0\}$, the model defined by Equations (61–64) is locally identified.

What is important about this theorem is that it highlights the importance of linearity in generating nonidentification. For a permutation of the linear-in-means model in the direction of any nonlinear function g, identification will hold. As nonlinearity seems to be a very standard feature of models with interactions, this result provides a relatively optimistic perspective on the identification problem, at least for the case of correctly specified models.

We believe that it should be relatively straightforward to extend the approach of McManus to show that identification is a generic property of nonlinear models with self-consistency constraints and are pursuing this in subsequent work.

3.6. Implications of self-selection for identification

Our discussion thus far has assumed that the rules by which individuals are sorted into groups has no implications for empirical analysis. Such an assumption implies that the group formation rule is independent of the determinants of individual choices and is thus unnatural in many contexts. Given the preferences we have assumed, one would expect individuals, when possible, to endogenously sort themselves, accounting for the effects of neighborhood characteristics and expected neighborhood behavior on payoff functions. Hence, there is the potential for self-selection bias. For decisions such as nonmarital births or dropping out of school, standard estimation methods may produce biased estimates due to the correlation of the $\epsilon_i\,(\omega_i)$'s with the determinants of sorting. To be clear, we do not explicitly account for equilibrium group formation, but rather approximate its effects through consideration of selection.

This issue has yet to be addressed in an extended fashion in the interactions literature. With reference to identification, what appears important is that self-selection may actually facilitate identification. Intuitively, self-selection can induce precisely the sort of nonlinearities or exclusion restrictions which generates identification in the earlier discussion.

To see this, we develop an example. Suppose that the econometric version of the linear-in-means model, Equation (55), describes the behavioral rule for all individuals in a population, but that we only observe those outcomes for individuals who have been sorted into neighborhoods in the sample. This can be justified by positing the existence of a reservation neighborhood for each individual. We assume that this means that there is a latent variable z_i which measures a family's evaluation of the neighborhood and such that a family is observed in neighborhood $n(i)$ if and only if $z_i > 0$. In turn, this latent variable can be written as

$$z_i = \gamma' \boldsymbol{R}_i + \eta_i, \tag{65}$$

where $\boldsymbol{R}_i$ is a vector of determinants of i's neighborhood evaluation. Finally, assume that the errors ϵ_i and η_i are zero mean, jointly normal with the variance/covariance matrix

$$\begin{bmatrix} \sigma_\epsilon^2 & \rho\sigma_\epsilon \\ \rho\sigma_\epsilon & 1 \end{bmatrix}, \tag{66}$$

and where $E(\epsilon_i \mid \boldsymbol{X}_i, \boldsymbol{Y}_{n(i)}, m^e_{n(i)}, \boldsymbol{R}_i) = E(\eta_i \mid \boldsymbol{X}_i, \boldsymbol{Y}_{n(i)}, m^e_{n(i)}, \boldsymbol{R}_i) = 0$.

This is precisely the model which is considered in Heckman (1979). Following his argument, since

$$E\left(\epsilon_i \mid z_i > 0\right) = \rho\sigma_\epsilon \lambda_i \left(\gamma' \boldsymbol{R}_i\right), \tag{67}$$

where, letting $\phi(\cdot)$ and $\Phi(\cdot)$ respectively denote the standard normal density and distribution,

$$\lambda\left(\gamma' \boldsymbol{R}_i\right) = \frac{\phi\left(\gamma' \boldsymbol{R}_i\right)}{\Phi\left(\gamma' \boldsymbol{R}_i\right)}, \tag{68}$$

a regression in which the model disturbance is orthogonal to the various regressors is

$$\omega_i = k + c' \boldsymbol{X}_i + d' \boldsymbol{Y}_{n(i)} + J m^e_{n(i)} + \rho\sigma_\epsilon \lambda\left(\gamma' \boldsymbol{R}_i\right) + \zeta_i. \tag{69}$$

What is important for our purposes is that the structure of this equation can facilitate identification. There are two distinct ways in which this can occur.

First, consider the case where each individual control is matched one-to-one with a contextual effect so that $E\left(\boldsymbol{X}_i \mid \boldsymbol{Y}_{n(i)}\right) = \boldsymbol{Y}_{n(i)}$. Assume as well that none of the variables in $\boldsymbol{R}_i$ are functionally dependent on $m^e_{n(i)}$, so that we may assume that the reduced form for $m^e_{n(i)}$ depends on $\boldsymbol{R}_i$. As discussed above, if $\rho\sigma_\epsilon = 0$, so there is no selection correction, this is Manski's (1993a,b) nonidentification example. However, in the presence of self-selection, the expected average choice within a neighborhood is, under self-consistency

$$m_{n(i)} = \frac{k}{1-J} + \left(\frac{1}{1-J}\right)\left(c' + d'\right)\boldsymbol{Y}_{n(i)} + \frac{\rho\sigma_\epsilon}{1-J}E\left(\lambda\left(\gamma'\boldsymbol{R}_i\right) \mid i \in n\left(i\right)\right), \tag{70}$$

so that a reduced form for individual behavior may be written as

$$\begin{aligned} \omega_i = {} & \frac{k}{1-J} + c'\boldsymbol{X}_i + \frac{1}{1-J}\left(Jc' + d'\right)\boldsymbol{Y}_{n(i)} + \rho\sigma_\epsilon\lambda\left(\gamma'\boldsymbol{R}_i\right) \\ & + \frac{J\rho\sigma_\epsilon}{1-J}E\left(\lambda\left(\gamma'\boldsymbol{R}_i\right) \mid i \in n\left(i\right)\right) + \zeta_i. \end{aligned} \tag{71}$$

In this reduced form regression, nonidentification when $\rho\sigma_\epsilon = 0$ follows immediately from observing that there are $2r + 1$ parameters and only $2r$ regressors. However, when there is a selection correction, two new regressors are introduced, $\lambda\left(\gamma'\boldsymbol{R}_i\right)$ and $E\left(\lambda\left(\gamma'\boldsymbol{R}_i\right) \mid i \in n\left(i\right)\right)$, but only one new parameter, $\rho\sigma_\epsilon$. This allows for identification so long as $\lambda\left(\gamma'\boldsymbol{R}_i\right)$ and $E\left(\lambda\left(\gamma'\boldsymbol{R}_i\right) \mid i \in n\left(i\right)\right)$ are not perfectly collinear, which requires that there is within-neighborhood variation in $\lambda\left(\gamma'\boldsymbol{R}_i\right)$. Notice that the nonlinearity of $\lambda(\cdot)$ ensures that the appearance of regressors in $\boldsymbol{R}_i$ which appear in either $\boldsymbol{X}_i$ or $\boldsymbol{Y}_{n(i)}$ does not imply nonidentification due to multicollinearity of the correction term with the other variables in the model.

This route to identification through selection correction is an example of the general identification condition stated in Theorem 6. In order to achieve identification, one needs an individual control whose neighborhood average is not a contextual effect. This is precisely what occurs when $\lambda\left(\gamma'\boldsymbol{R}_i\right)$ is introduced into the linear-in-means model, since $E\left(\lambda\left(\gamma'\boldsymbol{R}_i\right) \mid i \in n\left(i\right)\right)$ is not an element of the model even when selection is controlled for.

Second, identification may be achieved if $m_{n(i)}$ is a component of $\boldsymbol{R}_i$. Suppose that the expected average choice level is the only element in $\boldsymbol{R}_i$. The selection-corrected linear-in-means model is now

$$\omega_i = k + c'\boldsymbol{X}_i + d'\boldsymbol{Y}_{n(i)} + Jm^e_{n(i)} + \rho\sigma_\epsilon\lambda\left(\gamma m_{n(i)}\right) + \zeta_i. \tag{72}$$

The parameters in this regression will now be identified so long as the joint support of $\boldsymbol{X}_i$ and $\boldsymbol{Y}_{n(i)}$ does not lie in a proper linear subspace of $\boldsymbol{R}^{r+s}$ since the nonlinearity of the selection correction ensures that there is no linear dependence between $m^e_{n(i)}$ and the individual and neighborhood controls. Notice that this is the same reason for identification derived for the binary choice model; in both cases, the nonlinear dependence of $m^e_{n(i)}$ on $\boldsymbol{X}_i$ and $\boldsymbol{Y}_{n(i)}$ produces identification.

Of course, identifiability of model parameters does not say anything about the precision of the estimates facilitated by selection corrections. Intuitively, one will need substantial cross-neighborhood variation in $m^e_{n(i)}$ if the nonlinear dependence of the correction on this term is the basis for identification. Similarly, substantial variation in $\boldsymbol{R}_i$ will be needed if elements of this vector are highly correlated with combinations of $\boldsymbol{X}_i$ and $\boldsymbol{Y}_{n(i)}$, in order for the nonlinearity of the correction to avoid multicollinearity. Notice in this case, the presence of regressors in $\boldsymbol{R}_i$ which do not appear in $\boldsymbol{X}_i$ or $\boldsymbol{Y}_{n(i)}$ will likely prove valuable in practice.

To be clear, this discussion hardly exhausts the implications of selection corrections for identification. One issue concerns the relationship between the selection and behavior equations. There is no behavioral justification for the selection equation we have employed whereas ideally the selection equation will reflect individual optimization over a set of neighborhood choices and account for subsequent behavior which will occur in the neighborhood. Further, the analysis needs to be extended to cases where the joint normality of the selection and behavior disturbances is relaxed. Examples of nonparametric approaches to selection correction include Ahn and Powell (1993). What this example nevertheless demonstrates is that self-selection can, when accounted for, work to aid in identification, and hence clearly warrants further research.

3.7. Implications of multiple equilibria for identification

Finally, we observe that contrary to much of the conventional wisdom, the presence of multiple steady states can provide identification in and of itself, a possibility suggested in Manski (1993b, p. 539). To see this, suppose that all neighborhoods are composed of individuals with identical characteristics, so that $y_{n(i)} = \bar{y}$. Suppose that the J is greater than 1 and that $d\bar{y}$ is small enough relative to J that there are multiple steady states in a neighborhood. Finally, suppose that a fraction r of neighborhoods exhibit expected average choices consistent with the largest solution to $m = \tanh(d\bar{y} + Jm)$ and a fraction $1 - r$ exhibit average choices consistent with the smallest solution. In this case the determinant of the covariance matrix of $y_{n(i)}$ and $m_{n(i)}$ is $\bar{y}^2 Var\left(m_{n(i)}\right)$ which is nonzero unless $\bar{y}$ is zero. This would imply that in a regression of the form

$$\omega_i = k + d\boldsymbol{Y}_{n(i)} + Jm^e_{n(i)} + \epsilon_i, \tag{73}$$

J will be identified (although k and d of course will not be). The intuitive point is that variation in the realized equilibria across observations for a model with multiple equilibria can provide the leverage required to identify model parameters.

3.8. Dynamic models and rational expectations

Wallis (1980) provides an analysis of identification in classical rational expectations models that is closely related to the analysis of identification in the linear-in-means model. Suppose that the linear-in-means model is modified so that it now describes behavior at points in time, i.e.,

$$\omega_{i,t} = c'\boldsymbol{X}_{i,t} + d'\boldsymbol{Y}_{n(i),t} + Jm^e_{n(i),t} + \epsilon_{i,t}. \tag{74}$$

Let $\boldsymbol{\omega}_t$ denote the column vector of choices at t, X_t and Y_t denote matrices whose columns are the $\boldsymbol{X}_{i,t}$'s and $\boldsymbol{Y}_{n(i),t}$'s respectively, and C and D denote conformable matrices whose rows are always c' and d' respectively. Then a panel of observations on individuals can be written as

$$\boldsymbol{\omega}_t = C X_t + D Y_t + J \boldsymbol{m}_t^e + \epsilon_t. \tag{75}$$

When $D = 0$ and J is not a scalar, but rather a conformable matrix, one has the vector linear-in-means model version of the Wallis (1980) structural equation (2.1). These differences between the linear-in-means model and Wallis' model create new problems for identification. The identification problems which we have described in the linear-in-means model occur precisely because of the need to identify D. The identification problem is particularly acute when the Y matrix consists of neighborhood averages of X_t, as we have already seen.

Observing the connection between the linear-in-means model and Equation (75) suggests that fruitful connections exist between the literature we survey here and the classical rational expectations econometric work of Hansen and Sargent (1991) and Wallis (1980), among many others, as well as more recent work that extends that tradition to social interactions [Binder and Pesaran (1998a)] and to spatial rational expectations econometrics [Fingleton (1999)].

As Equations (74) and (75) make clear, the linear-in-means model is interpretable as a version of the Wallis model where $\boldsymbol{\omega}_t$ is scalar. The identification problem which occurs when $m_{n(i)}$ can be expressed as a linear combination of $\boldsymbol{Y}_{n(i)}$'s will occur in Wallis' model when $\boldsymbol{Y}_{n(i)} = E\left(\boldsymbol{X}_i \mid \boldsymbol{Y}_{n(i)}\right)$ for the various columns of Y_t. This connection between the problem of identification in the linear-in-means model which describes interactions in "space" with the problem of identification in linear rational expectations models in "time" suggests integrative future research along these lines should exist.

There are in fact many dimensions along which one can explore links between interactions-based models and rational expectations models. Hansen and Sargent (1991, p. 2), remark that

> "Work on rational expectations econometrics has divided into two complementary but differing lines. The first line aims more or less completely to characterize the restrictions that a model imposes on a vector stochastic process of observables, and to use those restrictions to guide efficient estimation. This line is a direct descendant of the full system approach to estimating simultaneous equation models . . .
>
> The second line of work is the application of method of moments estimators to estimating the parameters that appear in the Euler equations associated with dynamic optimization problems . . . "

Our discussion thus far has contrasted the linear-in-means model of social interactions with what Hansen and Sargent call the "first line" which is treated by Wallis in a framework particularly suited to comparison. In spatial optimization problems one could also develop a "second" line that parallels the Euler equation-based, methods of moments approach.

A key feature of dynamic rational expectations models is the potential for intertemporal interaction effects to influence identifiability. For example, suppose that individuals are affected by lagged group characteristics and lagged expected average behavior, so that

$$\omega_{i,t} = c'\boldsymbol{X}_{i,t} + d'\boldsymbol{Y}_{n(i),t-1} + Jm^{e}_{n(i),t-1} + \epsilon_{i,t}. \tag{76}$$

(We omit the constant k for expositional reasons.) In the case where all individual characteristics correspond one to one with neighborhood contextual effects (which is the Manski case of no identification in the linear-in-means model), this equation can be re-expressed as

$$m_{n(i),t} = c'\boldsymbol{Y}_{n(i),t} + d'\boldsymbol{Y}_{n(i),t-1} + Jm_{n(i),t-1}, \tag{77}$$

or

$$m_{n(i),t} = \left(\frac{1}{1-JL}\right) c'\boldsymbol{Y}_{n(i),t} + \left(\frac{1}{1-JL}\right) d'\boldsymbol{Y}_{n(i),t-1}. \tag{78}$$

where L denotes a lag operator. (We have assumed $|J| < 1$ so that the operator $1 - JL$ is invertible.) Substituting this expression into Equation (76),

$$\begin{aligned}\omega_{i,t} &= c'\boldsymbol{X}_{i,t} + d'\boldsymbol{Y}_{n(i),t-1} + J\left[\left(\frac{1}{1-JL}\right) c'\boldsymbol{Y}_{n(i),t-1} + \left(\frac{1}{1-JL}\right) d'\boldsymbol{Y}_{n(i),t-2}\right] + \epsilon_{i,t}\\ &= c'\boldsymbol{X}_{i,t} + \left(Jc' + d'\right)\boldsymbol{Y}_{n(i),t-1} + \left(\frac{J}{1-JL}\right)\left(Jc' + d'\right)\boldsymbol{Y}_{n(i),t-2} + \epsilon_{i,t},\end{aligned} \tag{79}$$

where the last line in the equation follows from $\left(\frac{1}{1-JL}\right) x_t = x_t + \left(\frac{J}{1-JL}\right) x_{t-1}$. Now, assume that the moment matrix generated by the elements of $(\boldsymbol{X}_{i,t}, \boldsymbol{Y}_{n(i),t-1}, \boldsymbol{Y}_{n(i),t-2}, \ldots)$ has full rank, so that the coefficient on each of the variables on the right hand side of Equation (79) is identified. Then the coefficients of the underlying structural model are also identified. To see this, observe first that c is identified by the coefficients on the regressors $\boldsymbol{X}_{i,t}$. J is identified because the coefficients on any corresponding elements of $\boldsymbol{Y}_{n(i),t-k}$ and $\boldsymbol{Y}_{n(i),t-k-1}$ with $k > 1$ proportional to J. Once c and J are identified, so is d from the coefficients on any set of regressors $\boldsymbol{Y}_{n(i),t-1}$. Intuitively, the timing of the interactions breaks the strict collinearity of the contextual and endogenous effects[4].

Finally, as an example of how the substantive economics in a dynamic model can influence identification, we consider a dynamic model of production complementarities of the type studied by Binder and Pesaran (1998a). In this model, the capital decisions

[4] Manski (1993b, p. 540) conjectures that a lagged linear-in-means model may be identified. Our verification of this conjecture suggests that the reason is not that the data are out of "temporal equilibrium" as Manski suggests, but rather that the collinearity of expected group outcomes and contextual effects is affected by dynamics in the interactions.

of a set of profit maximizing firms is studied. Each firm possesses a technology such that

$$Q_{i,t} = T_{i,t}K_{i,t}^{\alpha}, \tag{80}$$

where $T_{i,t}$ measures the level of firm i's technology and $K_{i,t}$ measures its capital stock. The rental price of capital for each firm is R_t. The level of technology of firm i is assumed to follow

$$T_{i,t} = A\exp\left(c'\log\boldsymbol{X}_{i,t} + d'E\left(\log\boldsymbol{X}_{i,t} \mid i \in n(i)\right) + JE\left(\log K_{i,t} \mid i \in n(i)\right) + \epsilon_{i,t}\right). \tag{81}$$

In this formulation, for any $\boldsymbol{w}$, $\log\boldsymbol{w}$ is the vector whose ith element is $\log w_i$. The shock $\epsilon_{i,t}$ is taken to be independent and identically distributed across both firms and time.

Firms are assumed to maximize the expectation of the present discounted value of their current and future profits. Each firm observes its own shock $\epsilon_{i,t}$ at the time it chooses $K_{i,t}$ but does not observe the shocks of other firms. Given our assumption that the technology shocks are independent across time, the discounted sum of profits breaks down into a sum of independent profit terms. Profit maximization with respect to the choice of firm-specific capital leads to the first-order condition

$$\begin{aligned}(1-\alpha)\log K_{i,t} &= \log(\alpha A) + c'\log(\boldsymbol{X}_{i,t}) + d'E\left(\log\boldsymbol{X}_{i,t} \mid i \in n(i)\right)\\ &\quad + JE\left(\log K_{i,t} \mid i \in n(i)\right) - \log R_t + \epsilon_{i,t}.\end{aligned} \tag{82}$$

Suppose that we have data for a cross-section of firms at fixed t. In this case, Equation (82) is an example of the linear-in-means model for which identification fails, since the group analog of each individual control appears in the structural equation; specifically, we have r individual controls $\log\boldsymbol{X}_{i,t}$ and r group level controls $E\left(\log\boldsymbol{X}_{i,t} \mid i \in n(i)\right)$, and the composite constant term $\log(\alpha) - \log R_t$, so by Theorem 6, the model is not identified[5].

Alternative routes to identification emerge when one allows for a richer dynamic structure to technology. Productivity spillovers generated by one firm onto another plausibly depend only on the current level of that firm's technology, not the particular path by which the firm arrived at that technology. On the other hand, the ability of any firm i to benefit from another firm's technology plausibly will depend on its own level of technology in the previous period as well as its characteristics today.

[5] Notice that we do not assume in this particular case that the averages of the individual characteristics $\log\boldsymbol{X}_{i,t}$ are known by members of a neighborhood. This has no effect on the analysis.

Formally, these assumptions on the dynamics of spillovers can be expressed in an equation for the logarithm of firm i's technology level such as

$$\begin{aligned} T_{i,t} = T_{i,t-1}^{\beta} A \exp\bigl[c' \log\left(\boldsymbol{X}_{i,t}\right) + d'E\left(\log \boldsymbol{X}_{i,t} \mid i \in n(i)\right) \\ + JE\left(\log K_{i,t} \mid i \in n(i)\right) + \epsilon_{i,t}\bigr]. \end{aligned} \tag{83}$$

The first order conditions for profit maximization now imply, after taking log's, that the capital level for each firm obeys

$$\begin{aligned} (1-\alpha)\log K_{i,t} = \beta \log T_{i,t-1} + \log(\alpha A) + c' \log(\boldsymbol{X}_{i,t}) \\ + d'E\left(\log \boldsymbol{X}_{i,t} \mid i \in n(i)\right) + JE\left(\log K_{i,t} \mid i \in n(i)\right) - \log R_t + \epsilon_{i,t}. \end{aligned} \tag{84}$$

What matters from the perspective of identification is that we now have an additional regressor $\log T_{i,t-1}$, whose group average does not appear in the equation. Hence for this model, we have $r+1$ individual effects, whereas we still have only r contextual effects. Hence, it will be possible to identify c, d, J. Of course, if this variable is not observable, it will itself have to be instrumented.

This example is only meant to be illustrative. Once one leaves the log linear framework, recent work, such as Pakes (1999) that treats nonlinearities seriously in firm dynamics, would be needed for the formulation of the stochastic processes characterizing firm behavior. Extending this kind of analysis to focus on measuring spillovers between firms seems to us to be a worthwhile area for future research. While production spillovers are the driving force behind the new growth theory, it is remarkable how little firm evidence of such spillovers actually exists. A primary reason for this is the weakness of the econometric methods which have been employed to obtain such evidence; Durlauf and Quah (1999) discuss many of the problems which exist with efforts to identify production externalities using cross-country growth data. Our belief is that the use of individual level data which follows interactions in the way we have described will yield much clearer inferences.

4. Further topics in identification

4.1. Panel data

The extension of identification results from cross-sections to panels is important for several reasons. First, panels will provide an opportunity for dealing with model misspecification which is not present in cross-sections. Hoffman and Plotnick (1996) is the only case we are aware of in which this argument is applied in an interactions context. Second, panels allow for intertemporal interactions which facilitate a richer notion of belief formation than we have used.

4.1.1. Fixed effects

To see how panel data can provide a way of dealing with misspecification, we start with the linear-in-means case. The panel analog to Equation (55) is

$$\omega_{i,t} = k + c'\boldsymbol{X}_{i,t} + d'\boldsymbol{Y}_{n(i,t),t} + Jm^e_{n(i,t),t} + \alpha_i + \epsilon_{i,t}. \tag{85}$$

In addition to introducing time subscripts on all variables, an unobservable fixed effect term α_i is now included.

In order to produce consistent estimates of c, d, and J, we follow the suggestion of Chamberlain (1984) and difference this equation with respect to t,

$$\Delta\omega_{i,t} = c'\Delta\boldsymbol{X}_{i,t} + d'\Delta\boldsymbol{Y}_{n(i,t),t} + J\Delta m^e_{n(i,t),t} + \Delta\epsilon_{i,t}. \tag{86}$$

Identification may now be treated in a fashion exactly analogous to that in Section 3, once first differences replace the levels used in the identification conditions. Notice that it is important to be careful about the assumption that regressors are orthogonal to errors in the differenced equation

$$E\left(\Delta\epsilon_{i,t} \mid \Delta\boldsymbol{X}_{i,t}, \Delta\boldsymbol{Y}_{n(i,t),t}, \Delta m^e_{n(i,t),t}\right) = 0,$$

since $\Delta\epsilon_{i,t}$ will not be white noise.

One implication of the panel data case with fixed effects is that variation in $\Delta m^e_{n(i,t),t}$ is useful in facilitating identification. In turn, this suggests that in environments which are slowly moving, J may be difficult to estimate precisely.

Analogous reasoning may be applied to the binary choice case. Suppose that the individual payoff function is now

$$V\left(\omega_{i,t}, \boldsymbol{Z}_{i,t}, \mu^e_{i,t}\left(\boldsymbol{\omega}_{-i,t}\right), \epsilon_{i,t}\left(\omega_{i,t}\right), \alpha_i\right), \tag{87}$$

where we again introduce time subscripts and a fixed effect α_i. We generalize our earlier development of the individual choice problem by assuming that the differential payoff between the two choices equals

$$c'\boldsymbol{X}_{i,t} + d'\boldsymbol{Y}_{n(i,t),t} + Jm^e_{n(i,t),t} + \epsilon_{i,t}(1) - \epsilon_{i,t}(-1) + \alpha_i. \tag{88}$$

So that the probability measure for $\omega_{i,t}$ obeys

$$\mu\left(\omega_{i,t} = 1 \mid \boldsymbol{X}_{i,t}, \boldsymbol{Y}_{n(i,t)}, m^e_{n(i,t)}\right) \sim \exp\left(\beta c'\boldsymbol{X}_{i,t} + \beta d'\boldsymbol{Y}_{n(i,t)} + \beta Jm^e_{n(i,t),t} + \alpha_i\right). \tag{89}$$

This equation is in a form which is estimable given methods derived in Honoré and Kyriazidou (1998). That paper also shows how one can adapt Manski's (1975, 1985) maximum score estimator so as to allow estimation of the model's parameters without assuming a logistic distribution for differences in the random utility terms.

With respect to the logistic case, Honoré and Kyriazidou (1998) show, extending an original insight of Chamberlain (1984), that if one has two consecutive observations on agent i, and if lagged $\omega_{i,t}$'s do not appear in the regressor matrix, then the conditional probability that either $\omega_{i,t-1} = -1$ or $\omega_{i,t} = 1$ given $\omega_{i,t-1} + \omega_{i,t} = 0$ is independent of α_i. This is the analogy to the differencing out of the fixed effect in the linear-in-means case. Honoré and Kyriazidou further show that if lagged $\omega_{i,t}$'s do appear in the regressor matrix, it is possible to modify their procedures and achieve identification so long as four consecutive observations on each agent are available. With respect to their conditions for identification, they appear to be consistent with the interactions-models we have been analyzing, once one allows for differenced rather than levels data.

This discussion has of course assumed that there is no self-selection into groups. Kyriazidou (1997a,b) provides conditions under which identification can occur for this case; extension of her methods to interactions-based environments would seem quite valuable.

4.1.2. Learning

An alternative use of panel data lies in the ability to model the expectations process as generated by learning. A simple way of doing this is to assume that the beliefs of an individual concerning the average choice in the population equals the realized average last period, so that social utility, for example, may be written as

$$S\left(\omega_{i,t}, \boldsymbol{X}_{i,t}, \boldsymbol{\omega}_{-i,t-1}\right) = -\sum_{j \neq i} \frac{J_{i,j}}{2}(\omega_{i,t} - \omega_{j,t-1})^2. \tag{90}$$

At this level, the dynamic interactions can be interpreted either as reflecting a primitive assumption either about individual preferences or concerning the way in which individuals form expectations of the contemporaneous behavior of others. This will not be so for more sophisticated learning models with interaction effects. Such models have been studied by Case (1992) and Munshi and Myaux (1998). In complementary work, Binder and Pesaran (1998a) provide an interesting analysis in which social interactions can be analyzed in a dynamic model with rational expectations.

4.2. Duration data

For contexts such as out-of-wedlock births or first sexual activity, it seems natural to consider interactions as they affect the probability of transition from one state to another; see Brewster (1994a,b) and Sucoff and Upchurch (1998) for empirical studies using this perspective. For such models, it is necessary to reformulate the nature of the interactions as they are manifested in a self-consistency condition analogous to Equation (16) in order to exploit the tools which have been developed to study duration data; these tools are well surveyed in Heckman and Singer (1984a, 1985) and Lancaster

(1990). To keep matters concrete, we will use the timing of out-of-wedlock births as an example.

Following standard notation [e.g., Amemiya (1985, Section 11.2), or Heckman and Singer (1985)], we let T denote the duration from $t = 0$ that an unmarried woman remains childless. The probability that this duration is less than any t, $\mu(T < t)$ is denoted as $F(t)$. For any interval δt, the probability that a childless woman at t becomes pregnant by $t + \delta t$ is

$$\mu(t \leqslant T < t + \delta t \mid T \geqslant t) = \frac{\mu(t \leqslant T < t + \delta t, T \geqslant t)}{\mu(T \geqslant t)}. \tag{91}$$

From this conditional probability, two standard functions of interest can be defined. First, the hazard function $\lambda(t)$ is defined as

$$\lambda(t) = \lim_{\delta t \Rightarrow 0} \frac{\mu(t \leqslant T < t + \delta t,\ T \geqslant t)}{\delta t \mu(T \geqslant t)} = \frac{F'(t)}{1 - F'(t)}. \tag{92}$$

Second, the survivor function $S(t)$ is defined as $1 - F(t)$. If $\lambda(t) = \lambda$, so that the hazard is independent of time, then the survivor function is

$$S(t) = \exp(-\lambda t), \tag{93}$$

which is the standard exponential form employed in many applications.

In order to make clear the basic identification issues we start with a baseline case. We assume that the time scale of the duration of interest is short relative to the time scale over which data are collected, so that all duration "spells" are completed. Formally, this assumption means that there is no "right censoring" of the data. This will not be appropriate for out-of-wedlock birth data, since of course not all unmarried females experience the event; nevertheless, the assumption is useful for exposition. Let t_i denote the time of first birth for individual i. If this timing is associated with probability density $f(\cdot)$, then the joint density for the I times is $\prod_i f(t_i)$. This joint probability will be determined by the individual hazards λ_i.

As before, we assume that the hazard for each individual under analysis depends on individual characteristics $\boldsymbol{X}_i$, neighborhood characteristics $\boldsymbol{Y}_{n(i)}$ and an expected neighborhood behavioral measure $m^e_{n(i)}$. In this context, $m^e_{n(i)}$ may be the expected value of either the within-neighborhood duration or the median group duration. We therefore assume that for each individual

$$\lambda_i = \lambda\left(\boldsymbol{X}_i, \boldsymbol{Y}_{n(i)}, m^e_{n(i)}\right), \tag{94}$$

so that the associated density for the duration is

$$f\left(t \mid \boldsymbol{X}_i, \boldsymbol{Y}_{n(i)}, m^e_{n(i)}\right) = \lambda\left(\boldsymbol{X}_i, \boldsymbol{Y}_{n(i)}, m^e_{n(i)}\right) \exp\left(-\lambda\left(\boldsymbol{X}_i, \boldsymbol{Y}_{n(i)}, m^e_{n(i)}\right) t\right). \tag{95}$$

The expected duration for individual i, conditional on these controls is

$$E\left(t \mid \boldsymbol{X}_i, \boldsymbol{Y}_{n(i)}, m^e_{n(i)}\right) = \lambda\left(\boldsymbol{X}_i, \boldsymbol{Y}_{n(i)}, m^e_{n(i)}\right)^{-1}, \tag{96}$$

and the median of the duration is given by the solution t^* to $F(t^*) = \frac{1}{2}$ which implies that

$$\tfrac{1}{2} = 1 - F(t^*) = \exp\left(-\lambda\left(\boldsymbol{X}_i, \boldsymbol{Y}_{n(i)}, m^e_{n(i)}\right) t^*\right), \tag{97}$$

which implies that t^* solves

$$\log 2 = \lambda\left(\boldsymbol{X}_i, \boldsymbol{Y}_{n(i)}, m^e_{n(i)}\right) t^*. \tag{98}$$

Equations (97) and (98) allow us to define self-consistent solutions for this model. Self-consistency with respect to expected duration times requires that

$$m^e_{n(i)} = m_{n(i)} = \int \lambda\left(\boldsymbol{X}_i, \boldsymbol{Y}_{n(i)}, m_{n(i)}\right)^{-1} \mathrm{d}F_{\boldsymbol{X}}, \tag{99}$$

where as before $F_{\boldsymbol{X}}$ is the probability distribution of characteristics within neighborhood $n(i)$. Similarly, self-consistency with respect to the neighborhood median requires that

$$m^e_{n(i)} = m_{n(i)} = \log 2 \int \lambda\left(\boldsymbol{X}_i, \boldsymbol{Y}_{n(i)}, m_{n(i)}\right)^{-1} \mathrm{d}F_{\boldsymbol{X}}. \tag{100}$$

These two expressions only differ by a constant of proportionality. Notice that we have assumed that each individual references on her entire neighborhood. It is possible to consider cases where the reference group is smaller, so that for example, one only references on individuals with similar individual characteristics. As we have already seen in the discussion of other models, the "width" of each individual's reference group plays a key role in identification.

We first consider identification in the parametric case under the assumption that the expected value of the duration time within a group is the relevant endogenous interaction. Following treatments such as Amemiya (1985, Section 11.2.3), we assume that the hazard function for individual i is exponential, so that

$$\lambda_i = \exp\left(c'\boldsymbol{X}_i + d'\boldsymbol{Y}_{n(i)} + Jm^e_{n(i)}\right). \tag{101}$$

We assume that $\boldsymbol{X}_i$ contains a constant term [Amemiya (1985, Equation 11.2.26)]. The associated likelihood function for the data will therefore be

$$L = \prod_i \exp\left(c'\boldsymbol{X}_i + d'\boldsymbol{Y}_{n(i)} + Jm^e_{n(i)}\right) \exp\left(-\exp\left(c'\boldsymbol{X}_i + d'\boldsymbol{Y}_{n(i)} + Jm^e_{n(i)}\right) t_i\right). \tag{102}$$

For this model, choosing parameter estimates for c, d, and J to maximize Equation (102) without imposing Equation (99) corresponds to the naive estimator we have described in the binary choice and linear-in-means cases.

Following the analysis in Amemiya (1985), identification in population requires that the expected value of the Hessian matrix of $\log L$ is nonsingular at the self-consistent solution (99). Letting $b = (c', d', J)'$ and $\boldsymbol{M}_i = (\boldsymbol{X}_i', \boldsymbol{Y}_{n(i)}', m_{n(i)}^e)'$, the expected value of the Hessian equals

$$E\left(\frac{\partial^2 \log L}{\partial b \partial b'} \mid \boldsymbol{M}_i\right) = -E\Bigl(\sum_i t_i \exp(b'\boldsymbol{M}_i)\boldsymbol{M}_i\boldsymbol{M}_i' \mid \boldsymbol{M}_i\Bigr). \tag{103}$$

Further, since $E(t_i \mid \boldsymbol{M}_i) = \lambda_i^{-1}$, it is further the case that

$$-E\Bigl(\sum_i t_i \exp(b'\boldsymbol{M}_i)\boldsymbol{M}_i\boldsymbol{M}_i' \mid \boldsymbol{M}_i\Bigr) = -\sum_i \boldsymbol{M}_i\boldsymbol{M}_i', \tag{104}$$

which means, dividing both sides by I, that identification asymptotically depends on the linear independence of the controls which constitute $\boldsymbol{M}_i$. This is the same condition which appeared in both the binary choice and linear-in-means. However, if $m_{n(i)}$ is the within-group mean, then by Equation (99) $m_{n(i)}$ is a nonlinear function of $\boldsymbol{Y}_{n(i)}$ and $F_{\boldsymbol{X}}$.

A nonlinear relationship of this type is the key condition for global identification in the binary choice model. Extensions of the analysis for binary choice may be made to the exponential hazard model as well as other parametric cases such as the Weibull, log normal and log-logistic distributions and can further be done for cases such as right censoring or for Cox's partial likelihood approach. A formal characterization of the conditions for identification in these various cases is left for future work. In particular, we believe that analogous conditions for local identification to those found in Theorem 7 can be developed for the current case.

4.3. Nonparametric approaches

4.3.1. Treatment effects

Our discussion of identification has assumed that a researcher possesses prior information concerning the form of the model under study, so that estimation occurs with respect to a finite set of parameters. In our context, this information has taken the form of both the functional form for individual behavior and, where selection into neighborhoods is an issue, the rules for neighborhood self-selection. Dissatisfaction with the assumption of such strong prior information has led to a vast literature on semi- and non-parametric approaches to estimation. In the context of interactions-based models, one can think of a nonparametric approach to estimation in the context of identifying a role for neighborhood characteristics on individual behavior while making relatively weak assumptions on the functional forms describing individual behavior. In turn, one can think of this question as analogous to the nonparametric identification of treatment effects, where group influences are the "treatment" whose effect we wish to uncover.

In this section, we develop approaches to both point and interval identification of interaction effects under substantially weaker modelling assumptions than we have employed thus far. First, following Heckman (1997), we show how the assumption that neighborhood interaction effects act as a "shifted outcome effect" combined with an exclusion restriction on the determinants of neighborhood membership, can lead to identification of an interaction effect. Second, following Manski (1995) and Manski and Pepper (1998), we show how one may relax this exclusion restriction and nevertheless obtain an upper bound on the interaction effect.

To make our analysis concrete, suppose that, following the work of Steinberg et al. (1996), we are interested in determining whether a peer group of "brains" (denoted as group 1) versus a peer group of "nonbrains" (denoted as group 0) affects individual student performance. Observations are available from G different schools, each of which contains students who are members of each such group. The variable $\xi_{i,g}$ tracks the group of individual i in school g. The goal of the exercise is to determine the effect of membership in group 1 versus group 0 on a continuous outcome variable $\omega_{i,g,\xi}$. Notice that we index according to both school and group. Membership in the brains groups is therefore our "treatment" and so we wish to measure the treatment effect. We let $\boldsymbol{X}_{i,g}$ denote those observable individual variables which directly determine $\omega_{i,g,\xi}$ and $\boldsymbol{R}_{i,g}$ denote those observable variables which determine whether i is a member of group 1. We refer to the average behaviors in the two groups as $m_{g,0}$ and $m_{g,1}$ with $\boldsymbol{m}_g = (m_{g,0}, m_{g,1})$. We have included g in the subscripts so that each observation refers to both an individual and the school which he attends.

For a given individual i, we assume that $\omega_{i,g,\xi}$ obeys

$$\omega_{i,g,\xi} = \phi\left(\xi_{i,g}, \boldsymbol{X}_{i,g}, \boldsymbol{R}_{i,g}, \boldsymbol{m}_g\right) + \epsilon_{i,g}\left(\xi_{i,g}\right), \tag{105}$$

for some function $\phi(\cdot,\cdot,\cdot,\cdot)$ where

$$E\left(\epsilon_{i,g}\left(\xi_{i,g}\right) \mid \xi_{i,g}, \boldsymbol{X}_{i,g}, \boldsymbol{R}_{i,g}, \boldsymbol{m}_g\right) = 0. \tag{106}$$

The identification question therefore refers to what can be learned about $\Delta_{i,g} = \omega_{i,g,1} - \omega_{i,g,0}$. Following Heckman (1997), one is typically interested in

$$E\left(\Delta_{i,g} \mid \boldsymbol{X}_{i,g}, \boldsymbol{R}_{i,g}, \boldsymbol{m}_g\right) = \phi\left(1, \boldsymbol{X}_{i,g}, \boldsymbol{R}_{i,g}, \boldsymbol{m}_g\right) - \phi\left(0, \boldsymbol{X}_{i,g}, \boldsymbol{R}_{i,g}, \boldsymbol{m}_g\right), \tag{107}$$

where the equality follows immediately from Equations (105) and (106). This is the expected value of the treatment for an individual with characteristics $\boldsymbol{X}_{i,g}$ and $\boldsymbol{R}_{i,g}$ in school g and represents the object which we wish to estimate. A distinct quantity of interest is $E\left(\Delta_{i,g} \mid \boldsymbol{X}_{i,g}, \boldsymbol{R}_{i,g}, \boldsymbol{m}_g, \xi_{i,g} = 1\right)$ which Heckman (1997) refers to as the effect of the "treatment on the treated for persons with characteristics" $\boldsymbol{X}_{i,g}$ and $\boldsymbol{R}_{i,g}$. Notice that the selection problem holds because there is information about $\epsilon_{i,g}(0)$ and $\epsilon_{i,g}(1)$ when the treatment, i.e., group membership, is a choice variable.

In order to identify $E\left(\Delta_{i,g} \mid \boldsymbol{X}_{i,g}, \boldsymbol{R}_{i,g}, \boldsymbol{m}_g\right)$, one proceeds as follows. First, assume that the effect of group membership is additive, so that

$$\omega_{i,g,1} - \omega_{i,g,0} = k\left(\boldsymbol{X}_{i,g}, \boldsymbol{m}_g\right), \tag{108}$$

for some function $k(\cdot,\cdot)$ which means that

$$k\left(\boldsymbol{X}_{i,g}, \boldsymbol{m}_g\right) = E\left(\Delta_{i,g} \mid \boldsymbol{X}_{i,g}, \boldsymbol{R}_{i,g}, \boldsymbol{m}_g\right). \tag{109}$$

Equation (108) is often referred to as a shifted outcome assumption. Notice that this is a minor generalization of Heckman (1997), although not Heckman and Robb (1985), in that we allow the k's to vary with respect to $\boldsymbol{X}_{i,g}$ and $\boldsymbol{m}_g$, which is natural if one thinks the treatment effect varies across individuals.

Next, consider what is estimable from the data. The group means $\boldsymbol{m}_g$ are of course observable. Further, one can estimate the conditional expectations of behavior for individuals given their group memberships, i.e.,

$$E\left(\omega_{i,g,0} \mid \boldsymbol{X}_{i,g}, \boldsymbol{R}_{i,g}, \boldsymbol{m}_g, \xi_{i,g} = 0\right), \tag{110}$$

and

$$E\left(\omega_{i,g,1} \mid \boldsymbol{X}_{i,g}, \boldsymbol{R}_{i,g}, \boldsymbol{m}_g, \xi_{i,g} = 1\right). \tag{111}$$

The identification of an endogenous interaction effect can be thought of as requiring that one can move from these conditional expectations to $E(\omega_{i,g,0} \mid \boldsymbol{X}_{i,g}, \boldsymbol{R}_{i,g}, \boldsymbol{m}_g)$ and $E(\omega_{i,g,1} \mid \boldsymbol{X}_{i,g}, \boldsymbol{R}_{i,g}, \boldsymbol{m}_g)$. To do this, it is necessary to be somewhat more careful about the process of group formation. We therefore assume that individuals join groups at least partially on the basis of the expected behavior in the groups, and that these expected behaviors are rational. This is nothing more than the self-consistency idea we have used throughout.

We now can consider the estimation of $k\left(\boldsymbol{X}_{i,g}, \boldsymbol{m}_g\right)$. Letting $\mu\left(\xi_{i,g} \mid \boldsymbol{X}_{i,g}, \boldsymbol{R}_{i,g}, \boldsymbol{m}_g\right)$ denote the conditional probability of group membership, it is immediate that

$$\begin{aligned} &E\left(\omega_{i,g,0} \mid \boldsymbol{X}_{i,g}, \boldsymbol{R}_{i,g}, \boldsymbol{m}_g\right) \\ &\quad = E\left(\omega_{i,g,0} \mid \boldsymbol{X}_{i,g}, \boldsymbol{R}_{i,g}, \boldsymbol{m}_g, \xi_{i,g} = 0\right) \mu\left(\xi_{i,g} = 0 \mid \boldsymbol{X}_{i,g}, \boldsymbol{R}_{i,g}, \boldsymbol{m}_g\right) \\ &\quad + E\left(\omega_{i,g,0} \mid \boldsymbol{X}_{i,g}, \boldsymbol{R}_{i,g}, \boldsymbol{m}_g, \xi_{i,g} = 1\right) \mu\left(\xi_{i,g} = 1 \mid \boldsymbol{X}_{i,g}, \boldsymbol{R}_{i,g}, \boldsymbol{m}_g\right), \end{aligned} \tag{112}$$

and

$$\begin{aligned} &E\left(\omega_{i,g,1} \mid \boldsymbol{X}_{i,g}, \boldsymbol{R}_{i,g}, \boldsymbol{m}_g\right) \\ &\quad = E\left(\omega_{i,g,1} \mid \boldsymbol{X}_{i,g}, \boldsymbol{R}_{i,g}, \boldsymbol{m}_g, \xi_{i,g} = 0\right) \mu\left(\xi_{i,g} = 0 \mid \boldsymbol{X}_{i,g}, \boldsymbol{R}_{i,g}, \boldsymbol{m}_g\right) \\ &\quad + E\left(\omega_{i,g,1} \mid \boldsymbol{X}_{i,g}, \boldsymbol{R}_{i,g}, \boldsymbol{m}_g, \xi_{i,g} = 1\right) \mu\left(\xi_{i,g} = 1 \mid \boldsymbol{X}_{i,g}, \boldsymbol{R}_{i,g}, \boldsymbol{m}_g\right). \end{aligned} \tag{113}$$

The right hand terms $E(\omega_{i,g,0} \mid \boldsymbol{X}_{i,g}, \boldsymbol{R}_{i,g}, \boldsymbol{m}_g, \xi_{i,g} = 1)$ and $E\ (\omega_{i,g,1} \mid \boldsymbol{X}_{i,g},\ \boldsymbol{R}_{i,g},\ \boldsymbol{m}_g, \xi_{i,g} = 0)$ are not observed since they refer to conditional expectations of behavior for individuals were they members of groups which they did not select into. Hence identification will only occur if some additional assumption overcomes this.

One such assumption is an exclusion restriction with respect to the variables which affect selection into groups versus variables which affect behavior once one is a member of a given group. Formally, we need the following. For every set of pairs $\boldsymbol{X}_{i,g}, \boldsymbol{R}_{i,g}$, and $\boldsymbol{X}_{i,g}, \boldsymbol{R}'_{i,g}$,

$$E\left(\omega_{i,g,0} \mid \boldsymbol{X}_{i,g}, \boldsymbol{R}_{i,g}, \boldsymbol{m}_g\right) = E\left(\omega_{i,g,0} \mid \boldsymbol{X}_{i,g}, \boldsymbol{R}'_{i,g}, \boldsymbol{m}_g\right), \tag{114}$$

and

$$E\left(\omega_{i,g,1} \mid \boldsymbol{X}_{i,g}, \boldsymbol{R}_{i,g}, \boldsymbol{m}_g\right) = E\left(\omega_{i,g,1} \mid \boldsymbol{X}_{i,g}, \boldsymbol{R}'_{i,g}, \boldsymbol{m}_g\right). \tag{115}$$

What this means is that there is a variable which affects selection but not expected behavior for each individual once that person is a group member.

Following Heckman (1997) and Manski (1995, p. 144), the shifted outcome restriction (108) and the exclusion restriction described by Equations (114) and (115) can be combined to conclude that

$$\begin{aligned}
&E\left(\omega_{i,g,1} \mid \boldsymbol{X}_{i,g}, \boldsymbol{R}_{i,g}, \boldsymbol{m}_g, \xi_{i,g} = 1\right) \mu\left(\xi_{i,g} = 1 \mid \boldsymbol{X}_{i,g}, \boldsymbol{R}_{i,g}, \boldsymbol{m}_g\right) \\
&\quad + E\left(\omega_{i,g,0} \mid \boldsymbol{X}_{i,g}, \boldsymbol{R}_{i,g}, \boldsymbol{m}_g, \xi_{i,g} = 0\right) \mu\left(\xi_{i,g} = 0 \mid \boldsymbol{X}_{i,g}, \boldsymbol{R}_{i,g}, \boldsymbol{m}_g\right) \\
&\qquad + k\left(\boldsymbol{X}_{i,g}, \boldsymbol{m}_g\right) \mu\left(\xi_{i,g} = 0 \mid \boldsymbol{X}_{i,g}, \boldsymbol{R}_{i,g}, \boldsymbol{m}_g\right) \\
&\quad = E\left(\omega_{i,g,1} \mid \boldsymbol{X}_{i,g}, \boldsymbol{R}'_{i,g}, \boldsymbol{m}_g, \xi_{i,g} = 1\right) \mu\left(\xi_{i,g} = 1 \mid \boldsymbol{X}_{i,g}, \boldsymbol{R}'_{i,g}, \boldsymbol{m}_g\right) \\
&\quad + E\left(\omega_{i,g,0} \mid \boldsymbol{X}_{i,g}, \boldsymbol{R}'_{i,g}, \boldsymbol{m}_g, \xi_{i,g} = 0\right) \mu\left(\xi_{i,g} = 0 \mid \boldsymbol{X}_{i,g}, \boldsymbol{R}'_{i,g}, \boldsymbol{m}_g\right) \\
&\qquad + k\left(\boldsymbol{X}_{i,g}, \boldsymbol{m}_g\right) \mu\left(\xi_{i,g} = 0 \mid \boldsymbol{X}_{i,g}, \boldsymbol{R}'_{i,g}, \boldsymbol{m}_g\right).
\end{aligned} \tag{116}$$

Other than $k\left(\boldsymbol{X}_{i,g}, \boldsymbol{m}_g\right)$, each of the terms in this expression can be estimated nonparametrically, and so $k\left(\boldsymbol{X}_{i,g}, \boldsymbol{m}_g\right)$ is identified.

Theorem 8. Nonparametric identification of endogenous interaction effect. *In the presence of self-consistent expectations, shifted outcomes of the form (108), and an exclusion restriction of the forms (114) and (115), the interaction effect is identified.*

Two features of this result are worth noting. First, under Theorem 8, one can estimate $E\left(k\left(\boldsymbol{X}_{i,g}, \boldsymbol{m}_g\right) \mid \boldsymbol{m}_g\right)$ and test for the average interaction effect in a population. Further, if one is willing to assume that

$$k\left(\boldsymbol{X}_{i,g}, \boldsymbol{m}_g\right) = J_g\left(m_{g,1} - m_{g,0}\right), \tag{117}$$

then the interaction parameter J_g for each school may be identified. In principle, cross-school variation in J_g could be employed to study the determinants of the strength of interactions.

Second, while Theorem 8 makes some progress in terms of relaxing the parametric assumptions of the interactions model, it is still strong in terms of the underlying behavioral assumptions. As stated by Heckman (1997, p. 449),

> "Any valid application of the method of instrumental variables for estimating these treatment effects in the case where the response to treatment varies among persons requires a behavioral assumption about how persons make their decisions about program participation. This issue cannot be settled by a statistical analysis."

In our context, the treatment is the membership in group 1 rather than group 0 and the instrument is characterized by Equations (114) and (115).

One approach to weakening the exclusion restriction on instruments is due to Manski and Pepper (1998). Following their analysis, we first replace our assumption of a shifted outcome variable, Equation (108) with

$$\omega_{i,g,1} - \omega_{i,g,0} = k\left(\boldsymbol{X}_{i,g}, \boldsymbol{m}_g\right). \tag{118}$$

This assumption means that the effect of shifting a person with individual characteristics $\boldsymbol{X}_{i,g}$ from group 0 to group 1 is bounded from above by $k\left(\boldsymbol{X}_{i,g}, \boldsymbol{m}_g\right)$. Second, we assume that a monotonic increase in the selection variables $\boldsymbol{R}_{i,g}$ never decreases the expected outcome for an individual within a given group. Formally, if $\boldsymbol{R}'_{i,g} \geqslant \boldsymbol{R}_{i,g}$, then

$$E\left(\omega_{i,g,\xi} \mid \boldsymbol{X}_{i,g}, \boldsymbol{R}_{i,g}, \boldsymbol{m}_g\right) \leqslant E\left(\omega_{i,g,\xi} \mid \boldsymbol{X}_{i,g}, \boldsymbol{R}'_{i,g}, \boldsymbol{m}_g\right), \quad \xi = 0, 1. \tag{119}$$

This assumption relaxes Equations (114) and (115) in that a monotonic increase from $\boldsymbol{R}_{i,g}$ to $\boldsymbol{R}'_{i,g}$ may have an effect on the conditional expectation of $\omega_{i,g,\xi}$, but this effect's sign must not be negative.

Under these assumptions, we have

$$\begin{aligned}
&E\left(\omega_{i,g,1} \mid \boldsymbol{X}_{i,g}, \boldsymbol{R}_{i,g}, \boldsymbol{m}_g, \xi_{i,g} = 1\right) \mu\left(\xi_{i,g} = 1 \mid \boldsymbol{X}_{i,g}, \boldsymbol{R}_{i,g}, \boldsymbol{m}_g\right) \\
&+ E\left(\omega_{i,g,0} \mid \boldsymbol{X}_{i,g}, \boldsymbol{R}_{i,g}, \boldsymbol{m}_g, \xi_{i,g} = 0\right) \mu\left(\xi_{i,g} = 0 \mid \boldsymbol{X}_{i,g}, \boldsymbol{R}_{i,g}, \boldsymbol{m}_g\right) \\
&\qquad + k\left(\boldsymbol{X}_{i,g}, \boldsymbol{m}_g\right) \mu\left(\xi_{i,g} = 0 \mid \boldsymbol{X}_{i,g}, \boldsymbol{R}_{i,g}, \boldsymbol{m}_g\right) \\
&\leqslant E\left(\omega_{i,g,1} \mid \boldsymbol{X}_{i,g}, \boldsymbol{R}'_{i,g}, \boldsymbol{m}_g, \xi_{i,g} = 1\right) \mu\left(\xi_{i,g} = 1 \mid \boldsymbol{X}_{i,g}, \boldsymbol{R}'_{i,g}, \boldsymbol{m}_g\right) \\
&+ E\left(\omega_{i,g,0} \mid \boldsymbol{X}_{i,g}, \boldsymbol{R}'_{i,g}, \boldsymbol{m}_g, \xi_{i,g} = 0\right) \mu\left(\xi_{i,g} = 0 \mid \boldsymbol{X}_{i,g}, \boldsymbol{R}'_{i,g}, \boldsymbol{m}_g\right) \\
&\qquad + k\left(\boldsymbol{X}_{i,g}, \boldsymbol{m}_g\right) \mu\left(\xi_{i,g} = 0 \mid \boldsymbol{X}_{i,g}, \boldsymbol{R}'_{i,g}, \boldsymbol{m}_g\right).
\end{aligned} \tag{120}$$

We may now consider the quantity, $Q\left(\boldsymbol{X}_{i,g}, \boldsymbol{R}_{i,g}, \boldsymbol{m}_g\right)$ defined as

$$\begin{aligned}
Q\left(\boldsymbol{X}_{i,g}, \boldsymbol{R}_{i,g}, \boldsymbol{m}_g\right) &= E\left(\omega_{i,g,1} \mid \boldsymbol{X}_{i,g}, \boldsymbol{R}_{i,g}, \boldsymbol{m}_g, \xi_{i,g} = 1\right) \mu\left(\xi_{i,g} = 1 \mid \boldsymbol{X}_{i,g}, \boldsymbol{R}_{i,g}, \boldsymbol{m}_g\right) \\
&+ E\left(\omega_{i,g,0} \mid \boldsymbol{X}_{i,g}, \boldsymbol{R}_{i,g}, \boldsymbol{m}_g, \xi_{i,g} = 0\right) \mu\left(\xi_{i,g} = 0 \mid \boldsymbol{X}_{i,g}, \boldsymbol{R}_{i,g}, \boldsymbol{m}_g\right).
\end{aligned} \tag{121}$$

This term is an observable analog of the expected outcome of an individual with observed characteristics $\boldsymbol{X}_{i,g}$, $\boldsymbol{R}_{i,g}$ and $\boldsymbol{m}_g$. Inequality (118) implies that

$$\begin{aligned}&Q\left(\boldsymbol{X}_{i,g},\boldsymbol{R}'_{i,g},\boldsymbol{m}_g\right)+k\left(\boldsymbol{X}_{i,g},\boldsymbol{m}_g\right)\mu\left(\xi_{i,g}=0\mid\boldsymbol{X}_{i,g},\boldsymbol{R}'_{i,g},\boldsymbol{m}_g\right)\\&\quad\geqslant Q\left(\boldsymbol{X}_{i,g},\boldsymbol{R}_{i,g},\boldsymbol{m}_g\right)+k\left(\boldsymbol{X}_{i,g},\boldsymbol{m}_g\right)\mu\left(\xi_{i,g}=0\mid\boldsymbol{X}_{i,g},\boldsymbol{R}_{i,g},\boldsymbol{m}_g\right),\end{aligned}\tag{122}$$

which may be rewritten as

$$\begin{aligned}&Q\left(\boldsymbol{X}_{i,g},\boldsymbol{R}'_{i,g},\boldsymbol{m}_g\right)-Q\left(\boldsymbol{X}_{i,g},\boldsymbol{R}_{i,g},\boldsymbol{m}_g\right)\\&\quad\geqslant k\left(\boldsymbol{X}_{i,g},\boldsymbol{m}_g\right)\left(\mu\left(\xi_{i,g}=0\mid\boldsymbol{X}_{i,g},\boldsymbol{R}_{i,g}\boldsymbol{m}_g\right)-\mu\left(\xi_{i,g}=0\mid\boldsymbol{X}_{i,g},\boldsymbol{R}'_{i,g},\boldsymbol{m}_g\right)\right).\end{aligned}\tag{123}$$

So long as

$$Q\left(\boldsymbol{X}_{i,g},\boldsymbol{R}'_{i,g},\boldsymbol{m}_g\right)-Q\left(\boldsymbol{X}_{i,g},\boldsymbol{R}_{i,g},\boldsymbol{m}_g\right)>0,\tag{124}$$

and

$$\mu\left(\xi_{i,g}=0\mid\boldsymbol{X}_{i,g},\boldsymbol{R}_{i,g},\boldsymbol{m}_g\right)-\mu\left(\xi_{i,g}=0\mid\boldsymbol{X}_{i,g},\boldsymbol{R}'_{i,g},\boldsymbol{m}_g\right)>0,\tag{125}$$

one can construct an upper bound on $k\left(\boldsymbol{X}_{i,g},\boldsymbol{m}_g\right)$. Formulating the bound using the fact that

$$\begin{aligned}&\mu\left(\xi_{i,g}=0\mid\boldsymbol{X}_{i,g},\boldsymbol{R}_{i,g},\boldsymbol{m}_g\right)-\mu\left(\xi_{i,g}=0\mid\boldsymbol{X}_{i,g},\boldsymbol{R}'_{i,g},\boldsymbol{m}_g\right)\\&\quad=\mu\left(\xi_{i,g}=1\mid\boldsymbol{X}_{i,g},\boldsymbol{R}'_{i,g},\boldsymbol{m}_g\right)-\mu\left(\xi_{i,g}=1\mid\boldsymbol{X}_{i,g},\boldsymbol{R}_{i,g},\boldsymbol{m}_g\right),\end{aligned}\tag{126}$$

we have Theorem 9.

Theorem 9. Construction of upper bound on the endogenous interaction effect. *Assume that Equations (118), (119), (124), and (125) hold. Then*

$$k\left(\boldsymbol{X}_{i,g},\boldsymbol{m}_g\right)\leqslant\frac{Q\left(\boldsymbol{X}_{i,g},\boldsymbol{R}'_{i,g},\boldsymbol{m}_g\right)-Q\left(\boldsymbol{X}_{i,g},\boldsymbol{R}_{i,g},\boldsymbol{m}_g\right)}{\mu\left(\xi_{i,g}=1\mid\boldsymbol{X}_{i,g},\boldsymbol{R}'_{i,g},\boldsymbol{m}_g\right)-\mu\left(\xi_{i,g}=1\mid\boldsymbol{X}_{i,g},\boldsymbol{R}_{i,g},\boldsymbol{m}_g\right)}.\tag{127}$$

A weakness of this result is that when it comes to interaction effects, it is probably more interesting to obtain a lower bound, since the presence of such effects is controversial. Notice, however, that the Manski and Pepper (1998) approach does suggest a way of constructing such a lower bound. In order to do so, one would need to find a variable which possesses the features that an increase (decrease) in its level would 1) both increase (decrease) the probability of selection into group 1, and 2) decrease (increase) the expected outcome for an individual conditional on the variable. Introspection suggests that it may be difficult to find such a variable, although there may be contexts where it holds.

In contrast, the assumption that one can find a variable in which both effects move in the same direction seems relatively plausible. For example, Manski and Pepper (1998) consider the question of how to bound the effect of the returns to schooling, using SAT as an instrument. It seems natural in this case to assume both that higher SAT's make additional schooling more likely and higher SAT's do not reduce the benefit of additional schooling if chosen.

The self-consistency conditions

$$m_{\xi,g} = \int E\left(\omega_{i,g,\xi} \mid \xi, \boldsymbol{X}, \boldsymbol{R}, \boldsymbol{m}_g\right) \mathrm{d}F_{\boldsymbol{X},\boldsymbol{R},g}, \quad \xi = 0, 1, \tag{128}$$

(where following our previous convention, $\mathrm{d}F_{\boldsymbol{X},\boldsymbol{R},g}$ denotes the joint distribution of $\boldsymbol{X}$ and $\boldsymbol{R}$ in school g), play an essential role in determining the quality of the bound in Equation (127). To see this, consider the extreme case where all individuals within a school g have identical values of $\boldsymbol{X}_{i,g}$ and $\boldsymbol{R}_{i,g}$, then the bound is undefined, since the numerator and denominator of Equation (125) will each equal 0. Alternatively, suppose that the distribution functions of individual characteristics are identical across schools, so $\mathrm{d}F_{\boldsymbol{X},\boldsymbol{R},g} = \mathrm{d}F_{\boldsymbol{X},\boldsymbol{R},g'}$. If the solutions to the self-consistency conditions (128) are unique, then this implies expected average outcomes must also be identical, i.e., $\boldsymbol{m}_g = \boldsymbol{m}_{g'}$. If $k\left(\boldsymbol{X}_{i,g}, \boldsymbol{m}_g\right) = k\left(\boldsymbol{m}_g\right)$, so that the bound does not depend on individual characteristics, then one can use cross-school information in that $k\left(\boldsymbol{m}_g\right)$ must be bounded by the inf of the upper bounds computed for each school in isolation. This type of argumentation seems a valuable area for future research.

At a minimum, this discussion illustrates two points. First, semi- and non-parametric approaches to inference can be adapted to achieve either point or interval identification of interaction effects. Second, the conditions required for identification require careful consideration of the underlying socioeconomic theories under analysis in order to identify appropriate instruments.

4.3.2. Duration data

Identification can also be considered for nonparametric approaches to duration data. As compellingly demonstrated by Heckman and Singer (1984b), errors in the assumed form of the hazard function and in the "mixing distribution" (by which they mean the distribution of unobservables) can lead to wildly misleading estimates. These problems of course are also relevant when interaction effects may be present. As far as we know, the extension of the methods studied by Heckman and Singer and subsequent authors to models with interactions has yet to be studied and it is beyond the scope of this paper to do so. However, we do sketch a slight extension to one approach to nonparametric identification in duration models, due to Elbers and Ridder (1982), in order to illustrate how such argumentation can in principle proceed; the reader is advised to see Heckman and Singer (1984c) for an evaluation of alternative conditions for nonparametric identification in this context.

Following Elbers and Ridder, suppose that the hazard function may be written as

$$\lambda_i(t) = a(t)\, h\,(\boldsymbol{M}_i)\upsilon_i. \tag{129}$$

Relative to our original treatment of hazards, this incorporates two additional terms: $a(t)$ which allows for duration dependence, and υ_i which allows for unobserved heterogeneity. The υ_i's are assumed to be drawn from a common distribution $F_\upsilon(\cdot)$ with associated density $f_\upsilon(\cdot)$. Elbers and Ridder (1982) show that subject to appropriate regularity conditions, if $F\,(t \mid \boldsymbol{M}_i)$ is nondefective (which means that all spells are completed), then it is possible to identify $a(\cdot)$, $h(\cdot)$, and $f_\upsilon(\cdot)$. They do this as follows.

Define the conditional survivor function for individual i as

$$S\,(t, \boldsymbol{M}_i, \upsilon_i) = \exp\left(\int_o^t a(r)\, h\,(\boldsymbol{M}_i)\, \upsilon_i \mathrm{d}r\right) = \exp\,(A(t)\, h\,(\boldsymbol{M}_i)\, \upsilon_i)\,, \tag{130}$$

where $A(t) = \int_o^t a\,(r)\,\mathrm{d}r$ and the conditional survivor function

$$S\,(t, \boldsymbol{M}_i) = \int S\,(t, \boldsymbol{M}_i, \upsilon)\,\mathrm{d}F_\upsilon = \int \exp\,(\tau\upsilon)\,\mathrm{d}F_\upsilon, \tag{131}$$

where $\tau = A(t)h\,(\boldsymbol{M}_i)$. The last term in Equation (131) indicates how the conditional survivor function is the LaPlace transform of $\mathrm{d}F_\upsilon$. The analyst is assumed to observe a family of nondefective distribution functions $G\,(t, \boldsymbol{M}_i) = 1 - S\,(t, \boldsymbol{M}_i)$ from which he wishes to recover $a(\cdot)$, $h(\cdot)$, and $f_\upsilon(\cdot)$. Elbers and Ridder (1982) assume that 1) υ is nonnegative with mean 1, 2) $\boldsymbol{M}_i$ lies in an open set in the k-dimensional reals for some k, and 3) $h(\cdot)$ is defined on this open set and is non-negative, differentiable, and nonconstant on the set.

Working through the proof that these conditions allow for identification, reveals the following. First, if one differentiates $G\,(t, \boldsymbol{M}_i)$ with respect to t, $h\,(\boldsymbol{M}_i)$ may be recovered regardless of whether a self-consistency condition like Equation (99) holds when $m_{n(i)}$ is a component of $\boldsymbol{M}_i$. Second, Elbers and Ridder (1982) exploit the LaPlace transform relationship to obtain a differential equation by assuming, without loss of generality, that $\boldsymbol{M}_i$ is one-dimensional. This argument requires differentiability and nonconstancy of $h\,(\boldsymbol{M}_i)$.

In order to generalize this step to allow for endogenous interactions with a self-consistency condition such as Equation (99), the reference group for each individual i must be broad enough to allow differentiability with respect to a nontrivial subvector of $\boldsymbol{M}_i$. Specifically, one needs to be able to vary $\boldsymbol{X}_i$ and $\boldsymbol{Y}_{n(i)}$ without $m_{n(i)}$ varying. For example, suppose that within each neighborhood, all $\boldsymbol{X}_i$'s are identical. In this case, the self-consistent average choice level is

$$m_{n(i)} = E\,\left(t \mid \boldsymbol{X}_i, \boldsymbol{Y}_{n(i)}, m_{n(i)}\right) = \int t\mathrm{d}G\,\left(t, \boldsymbol{X}_i, \boldsymbol{Y}_{n(i)}, m_{n(i)}\right). \tag{132}$$

Generically, Equation (132) will have only a finite number of self-consistent solutions (if such solutions exist). Therefore, in this case $\left(\boldsymbol{X}_i, \boldsymbol{Y}_{n(i)}\right)$ cannot be varied independently of $m_{n(i)}$ and it is not obvious how to adapt the Elbers and Ridder (1982, p. 405) proof to this case. Hence for the case of arbitrarily fine-grained reference groups, identification is currently problematic.

In contrast, suppose that individuals reference on a coarse group G. In this case the self-consistency condition is

$$m_{n(i)} = \int_X E\left(t \mid X, \boldsymbol{Y}_{n(i)}, m_{n(i)}\right) \mathrm{d}F_X = H\left(m_{n(i)}\right), \tag{133}$$

where $\mathrm{d}F_X$ is the distribution of X's within $n(i)$. In this case, it is possible to locate a nontrivial set of sufficient conditions on $\mathrm{d}G\left(t, \boldsymbol{X}_i, \boldsymbol{Y}_{n(i)}, m_{n(i)}\right)$ such that the hazard function is differentiable with respect to $\left(\boldsymbol{X}_i, \boldsymbol{Y}_{n(i)}\right)$ on the self-consistent solutions defined by Equation (133). This appears to be sufficient to extend Elbers and Ridder's identification argument to the case of interactions.

5. Sampling properties

In this section, we develop some asymptotics for the parameter estimates for interactions-based models and consider the effects on such estimates of omitted variables. The sampling properties for data generated by interactions-based models are no different from that associated with standard discrete choice and linear regression models. The critical property which one needs to verify is that the behavioral data obey the standard limits theorem necessary for asymptotics when there is sufficient dependence across observations to induce multiple equilibria. Similarly, the effects of omitted variables mirror results found in other contexts. We therefore focus on the binary choice case.

5.1. Laws of large numbers

Despite the dependence introduced by interactions, the data generated by the noncooperative version of the binary choice model with interactions generates a law of large numbers. Brock and Durlauf (1995) showed this for the special case where $h_i = h$ $\forall\, i$; it is straightforward [cf., Ash (1972, p. 234)] to extend this result to non-identically distributed choices.

Theorem 10. Law of large numbers for realized average choice levels in noncooperative version of the binary choice model with interactions. *Suppose that a population of agents holds a common belief that the expected value of the average population choice is m^*, where m^* is a solution to Equation (22). Then a weak law of large numbers holds: I becomes arbitrarily large such that we have*

$$\lim \bar{\omega}_I \Rightarrow_w m^*. \tag{134}$$

5.2. Naive estimator

The "naive" estimator whose identification properties we have analyzed does not introduce any new econometric issues with respect to asymptotic normality. Theorem 11 is standard; necessary conditions for it are given, for example, in McFadden (1984, p. 1399). The specific conditions we cite are found in Amemiya (1985, p. 270)[6].

Theorem 11. Consistency and asymptotic normality of naive estimates in the binary choice model with global interactions. *Let* $b = (k, c, d, J)'$ *and* $\boldsymbol{M}_i = (1, \boldsymbol{X}_i', \boldsymbol{Y}_{n(i)}', m_{n(i)}^e)'$. *If the binary choice model with interactions is globally identified, and if*

i. b lies in an open, bounded subset of R^{r+s+2}.

ii. $\lim_{I \Rightarrow \infty} I^{-1} \sum_{i \in I} \boldsymbol{M}_i \boldsymbol{M}_i'$ *is a finite nonsingular matrix.*

iii. The empirical distribution function of $\boldsymbol{M}_i$ *converges to a distribution function.*

Then, the maximum likelihood estimates $\hat{b}_I$ *of the binary choice model with global interactions are consistent and asymptotically normal with limiting behavior*

$$I^{1/2}(\hat{b}_I - b) \Rightarrow_w N(0, \vartheta^{-1}), \tag{135}$$

where

$$\vartheta = \lim_{I \Rightarrow \infty} I^{-1} \sum_i \frac{\exp(\hat{b}'\boldsymbol{M}_i)}{(1 + \exp(\hat{b}'\boldsymbol{M}_i))^2} \boldsymbol{M}_i \boldsymbol{M}_i'. \tag{136}$$

(ϑ is of course the suitably normalized information matrix of the likelihood function and is consistently estimable for this model.)

5.3. Asymptotics for data generated by social planner

Models which incorporate realized contemporaneous interactions between individuals introduce several mathematical complexities relative to standard econometric models. As noted before, this occurs because of the quadratic terms which appear in the likelihood. The following theorem is proved in the Appendix; unlike Theorem 10 it does not apply to the case of heterogeneous h_i's, although results in Amaro de Matos and Perez (1991) suggest this can be done.

Theorem 12. Large economy limit for realized average choice levels in social planner's version of the binary choice model with interactions. *Suppose that the vector of choices in a population is determined by a social planner with preferences*

[6] In Amemiya (1985), it is also assumed that the $\boldsymbol{M}_i$ elements are uniformly bounded when asymptotic normality is proved, which contradicts our identification assumption that the $\boldsymbol{Y}_{n(i)}$'s are unbounded. However, as Amemiya points out, the boundedness assumption can be dispensed with.

consistent with Equation (44). The sample mean of these choices converges weakly, that is,

$$\lim \bar{\omega}_I \Rightarrow_w m^*, \tag{137}$$

where m^* is the solution to $m^* = \tanh(\beta h + \beta J m^*)$ with the same sign as h.

Unfortunately, maximum likelihood estimation has yet to be developed for data generated by a social planner problem of the type we have studied. While techniques developed in Amaro de Matos and Perez (1991), Brock (1993), and Ellis (1985) all suggest that the development of these asymptotics is feasible, the argument seems sufficiently complicated that we are not comfortable making a conjecture on the asymptotic distribution of the estimator. In the Appendix, we provide some initial discussion of these issues to illustrate how such a theory could be developed.

5.4. Unobserved variables

Perhaps the most serious criticism made of efforts to identify interaction effects is the difficulty in identifying interaction effects in the presence of unobserved individual or group characteristics. This is true because the main groupings for which interactions are conjectured to exist, neighborhoods, schools, firms, etc., are endogenously determined. Presumably, neighborhood contextual and endogenous characteristics influence individual choices as to neighborhood membership. Hence, it seems very likely that omitted variables which influence individual behavior once that person is a member of a neighborhood will also be correlated with the various group effects which are captured in a statistical model. This point is distinct from the self-selection issues which are discussed above.

In particular, we are interested in determining how omitted variables will affect inferences concerning J. We do this following a maximum likelihood approach due to Cameron and Heckman (1998). This approach is straightforward to describe in sample, rather than population terms which is why we place it here.

In our framework, assume that the binary choices ω_i are coded 0, 1 and are generated by the probability model

$$\begin{aligned}\mu&\left(\omega_i = 1 \mid \boldsymbol{X}_{i,\mathrm{o}}, \boldsymbol{Y}_{n(i),\,\mathrm{o}}, \boldsymbol{X}_{i,\mathrm{u}}, \boldsymbol{Y}_{n(i),\,\mathrm{u}}, m^e_{n(i)}\right)\\ &= F_\epsilon\left(k + c'_{\mathrm{o}}\boldsymbol{X}_{i,\mathrm{o}} + d'_{\mathrm{o}}\boldsymbol{Y}_{n(i),\,\mathrm{o}} + c'_{\mathrm{u}}\boldsymbol{X}_{i,\mathrm{u}} + d'_{\mathrm{u}}\boldsymbol{Y}_{n(i),\,\mathrm{u}} + Jm^e_{n(i)}\right),\end{aligned} \tag{138}$$

where subscripts o and u refer to observed and unobserved variables, respectively. We assume that F_ϵ is the logistic distribution.

Let $\Theta' = (k, c_{\mathrm{o}}', d_{\mathrm{o}}', J)$, $\boldsymbol{Z}_i' = (1, \boldsymbol{X}_{i,\mathrm{o}}, \boldsymbol{Y}_{n(i),\mathrm{o}}, m_{n(i)}^e)$ and $\eta_i = c_{\mathrm{u}}'\boldsymbol{X}_{i,\mathrm{u}} + d_{\mathrm{u}}'\boldsymbol{Y}_{n(i),\mathrm{u}}$. This means the true probability structure can be rewritten as

$$\mu\left(\omega_i = 1 \mid \boldsymbol{X}_{i,\mathrm{o}}, \boldsymbol{Y}_{n(i),\mathrm{o}}, \boldsymbol{X}_{i,\mathrm{u}}, \boldsymbol{Y}_{n(i),\mathrm{u}}, m_{n(i)}^e\right) = F_\epsilon\left(\Theta'\boldsymbol{Z}_i + \tau\eta_i\right), \tag{139}$$

which produces a likelihood function of the form

$$L = I^{-1}\sum_i \left(\omega_i \log F_\epsilon\left(\Theta'\boldsymbol{Z}_i + \tau\eta_i\right) + (1-\omega_i)\log\left(1 - F_\epsilon\left(\Theta'\boldsymbol{Z}_i + \tau\eta_i\right)\right)\right). \tag{140}$$

The likelihood function is concave in Θ. The derivatives of the likelihood function (140) may be written as

$$L_\Theta = I^{-1}\sum_i \boldsymbol{Z}_i\left(\omega_i - F_\epsilon\left(\Theta'\boldsymbol{Z}_i + \tau\eta_i\right)\right), \tag{141}$$

$$L_{\Theta,\Theta} = -I^{-1}\sum_i f_\epsilon\left(\Theta'\boldsymbol{Z}_i + \tau\eta_i\right)\boldsymbol{Z}_i\boldsymbol{Z}_i', \tag{142}$$

$$L_{\Theta,\tau} = -I^{-1}\sum_i f_\epsilon\left(\Theta'\boldsymbol{Z}_i + \tau\eta_i\right)\boldsymbol{Z}_i\eta_i, \tag{143}$$

where $\tau = 0$ is the case where there are no unobservables and $\tau = 1$ is the case where there are unobservables. The maximum likelihood estimate of Θ must obey

$$L_\Theta\left(\Theta(\tau),\, \tau\right) = 0, \tag{144}$$

where we have written the likelihood as a function of the unknown parameters Θ and have allowed the estimate of Θ to depend on τ. Further,

$$L_{\Theta,\Theta}\left(\Theta(\tau), \tau\right)\frac{\mathrm{d}\Theta}{\mathrm{d}\tau} + L_{\Theta,\tau}\left(\Theta(\tau), \tau\right) = 0, \tag{145}$$

which implies that

$$\frac{\mathrm{d}\Theta}{\mathrm{d}\tau} = -L_{\Theta,\Theta}\left(\Theta(\tau), \tau\right)^{-1} L_{\Theta,\tau}\left(\Theta(\tau), \tau\right). \tag{146}$$

Integrating both sides of this expression produces

$$\Theta(0) - \Theta(1) = \int_0^1 L_{\Theta,\Theta}\left(\Theta(\tau), \tau\right)^{-1} L_{\Theta,\tau}\left(\Theta(\tau), \tau\right)\mathrm{d}\tau. \tag{147}$$

This difference describes the effect of misspecification since, as noted above, $\tau = 0$ corresponds to the case of no unobservables whereas $\tau = 1$ corresponds to the case with unobservables as we have formulated them.

In general, one cannot determine the sign of $\Theta(0) - \Theta(1)$; this is not surprising since it is known in contexts such as measurement error that unambiguous statements about directions of bias cannot be made. However, as recognized by Bretagnolle and Huber-Carol (1988), one can determine the sign of this bias in special cases. For example, if the elements of $L_{\Theta,\Theta}$ are all negative and the elements of $L_{\Theta,\tau}$ are all positive, then the coefficients in the misspecified model are all biased upwards.

Using the formula (147), one can compute the bias associated with the parameter J. For the case where the vector $\boldsymbol{X}_i$ is replaced with a scalar x_i the vector $\boldsymbol{Y}_{n(i)}$ is replaced with a scalar $y_{n(i)}$, one can compute

$$\begin{aligned} J(0) - J(1) = &\int_0^1 \left(\left(L^{-1}_{\Theta,\Theta} \right)_{4,1} \left(I^{-1} \sum_i f_\epsilon \left(\Theta' \boldsymbol{Z}_i + \tau\eta_i \right) \eta_i \right) \right) \mathrm{d}\tau \\ &+ \int_0^1 \left(\left(L^{-1}_{\Theta,\Theta} \right)_{4,2} \left(I^{-1} \sum_i f_\epsilon \left(\Theta' \boldsymbol{Z}_i + \tau\eta_i \right) x_i \eta_i \right) \right) \mathrm{d}\tau \\ &+ \int_0^1 \left(\left(L^{-1}_{\Theta,\Theta} \right)_{4,3} \left(I^{-1} \sum_i f_\epsilon \left(\Theta' \boldsymbol{Z}_i + \tau\eta_i \right) y_{n(i)} \eta_i \right) \right) \mathrm{d}\tau \\ &+ \int_0^1 \left(\left(L^{-1}_{\Theta,\Theta} \right)_{4,4} \left(I^{-1} \sum_i f_\epsilon \left(\Theta' \boldsymbol{Z}_i + \tau\eta_i \right) m_{n(i)} \eta_i \right) \right) \mathrm{d}\tau, \end{aligned} \tag{148}$$

where $\left(L^{-1}_{\Theta,\Theta} \right)_{i,j}$ is the i,jth element of $L^{-1}_{\Theta,\Theta}$ and we impose self-consistency of beliefs (i.e., substituting $m_{n(i)}$ for $m^e_{n(i)}$). The term $\left(L^{-1}_{\Theta,\Theta} \right)_{4,4}$ is nonpositive whereas the other inverse elements of these integrals are of ambiguous sign. This is the only sense in which one might say there is a presumption that estimates of interaction effects are biased towards finding them because of omitted variables.

Cameron and Heckman (1998) show how the Heckman and Singer (1984b) nonparametric likelihood estimator can be used to estimate the distribution of the unobserved η_i's and thereby compute unbiased estimates of the parameters of the observables. They make a compelling argument that the production of "heterogeneity-corrected estimates" is essential in conducting assessments of policy experiments. We are currently pursuing the development of this idea to produce estimates of interaction effects which are robust to omitted individual and group characteristics.

6. Statistical analysis with grouped data

In this section, we explore some of the approaches to identifying interactions which have been developed for aggregated data. Our discussion so far has assumed that individual level observations are available to the researcher. In contexts such as economic growth or crime rates, it is often the case that only group-level data is available. As a result, there has been a distinct literature which deals with uncovering interactions from aggregated data series.

6.1. Differences in cross-group behavior

One approach to the identification of interactions from group level data is due to Glaeser et al. (1996) and extended in Glaeser and Scheinkman (1998). The basic insight of this work focuses on the implications of interactions for the distribution of cross-group differences in choices. Consider a collection of groups, $N_1 \ldots N_I$, each of which has n members. In each of the groups, individuals face a binary choice. If the individuals within each group are identical, and their choices are independent of one another, then sample means of choices within each group will scale according to the law of large numbers. Supposing that the probability of choosing 1 is p, then the variance of the sample average for each of the two groups is $n^{-1}p(1-p)$. This means that the cross-group variance converges weakly to zero at rate n^{-1}. Observations that the cross-group variance scales at a slower rate, i.e., that the cross-group differences in average choice vary too much to be consistent with the sample variance under the null hypothesis of independent and identically distributed choices, is taken as evidence of social interactions.

Glaeser et al. (1996) apply their analysis to the study of cross-city crime rates. They find that even after controlling for city-specific socioeconomic variables, there are cross-city differences in crime rates which are far greater than would be consistent with individuals making independent choices within cities and conclude that this evidence is strongly supportive of an interactions approach.

6.2. Spatial patterns

Topa (1997) has attempted to identify and measure interactions through the use of spatial data. The basic idea of this work is to take seriously the idea that geographic proximity is a proxy for social proximity. Topa does this by considering the relationship between unemployment rates in census tracts in Chicago. Since census tracts typically vary in size between 2000 to 8000 residents, these units would seem to be good proxies for neighborhoods. Topa further assumes that the social distance between any two adjacent tracts is 1, the social distance between a tract and another tract that can be reached by travelling through a single other tract is 2, etc. Using these assumptions, he formulates the determinants of unemployment in a given tract n at time t, $\bar{\omega}_{n,t}$, as

$$\bar{\omega}_{n,t} = \varphi\left(c'\bar{\boldsymbol{X}}_{n,t} + J'\bar{\boldsymbol{\omega}}_{n,D,t} + \epsilon_{i,t}\right), \tag{149}$$

where $\bar{\boldsymbol{X}}_{n,t}$ denotes census tract averages of a set of individual characteristics, $\bar{\boldsymbol{\omega}}_{n,D,t}$ is a vector of average unemployment rates for tracts at social distances $1,2,\ldots D$ away from tract n and $\varphi(\cdot)$ is a nonlinear function generated by the stochastic model (a contact process) used to motivate the econometrics. Topa estimates this model using indirect inference methods and finds evidence of interaction effects in the sense of a statistically significant J vector.

Conley and Topa (1999) extend this analysis by attempting to identify what role different measures of distance play in explaining these spatial correlations. In particular, they construct measures of neighborhood distance based on 1) physical distance, 2) travel time, 3) ethnicity, and 4) occupation. Their results suggest that physical and occupational distance explain residual correlations across unemployment rates within census tracts once intra-tract characteristics have been controlled for. Akerlof (1997) demonstrates the theoretical importance of integrating social distance into economic analysis and provides a range of interesting potential applications; the Conley and Topa work should help produce empirical measures of various types of social distance.

6.3. Ecological inference

In the political science literature, there have been some efforts to identify group effects under the rubric of what is known as the ecological inference problem. In the basic version of this problem, a researcher possesses data on the number of whites and African Americans in each of a set of I neighborhoods, as well as the number of votes received by a white and an African American candidate in the same neighborhoods. The researcher's goal is to determine the relationship between racial composition of a neighborhood and the distribution of votes by race. Since the researcher is attempting to infer individual behavior from aggregate statistics, the inference is referred to as "ecological".

Ecological inference has generated a literature which has recently begun to grow [Goodman (1953), Freedman et al. (1991, 1998), King (1997)]. We follow the exposition of Freedman et al. (1998). Letting r_i denote the percentage of African Americans in a neighborhood and υ_i as the percentage of votes accrued by an African American candidate, the standard ecological regression is

$$\upsilon_i = pr_i + q(1 - r_i) + \epsilon_i, \tag{150}$$

where p is the probability with which African Americans vote for an African American candidate and q is the probability with which whites vote for an African American candidate. This equation is estimable by ordinary least squares. Alternative approaches to ecological inference typically modify this equation. For example, King (1997) proposes treating the racial voting propensities as neighborhood-specific draws from a common distribution rather than as constants.

From the perspective of the sorts of data sets and models of interest to economists, we suspect that ecological inference as it has been developed is of limited interest. The formulation of the regressions fails to correspond in a natural way to the aggregate of individual decisions into group behavioral percentages in a way consistent with a choice-theoretic framework. Indeed, the consistency of aggregated voting behavior with more than one behavioral model is precisely the basis on which Freedman et al. (1991) argued that evidence of differences in p and q could not be interpreted in terms of underlying differences in behavior between white and African American voters. As far as we know, no one has yet shown that the statistical tools in the ecological inference literature can complement other techniques for the recovery of socioeconomic structure. However, Cross and Manski (1999) suggest new directions along these lines which may both clarify what structural mechanisms can be revealed by aggregate data as well as show how ecological regression relates to omitted variables problems in econometrics.

7. Evidence

In this section, we survey some of the evidence which has been adduced to detect the presence of and to measure the magnitude of interactions. We divide the empirical literature into two parts. The first part assumes that the regression of individual outcomes on individual and group level variables represents a correctly specified model. In particular, the analysis assumes that there are no omitted variables which will generate coefficient inconsistency. The second approach accounts for the possibility of such omitted variables and explores ways to correct for inconsistency either through choice of data sets or econometric techniques.

7.1. Analyses under assumption of correct specification

In this subsection we review some prominent empirical analyses of interactions.

7.1.1. Neighborhood effects in youth and adult outcomes

Perhaps the most widely empirically studied area of interactions concerns the effects on adults of the neighborhoods in which they grew up. The typical analysis of this type computes a regression of the form

$$\omega_{i,t+1} = a + c'\boldsymbol{X}_{i,t} + d'\boldsymbol{Y}_{n(i),t} + \epsilon_{i,t+1}, \tag{151}$$

where, as before individual family characteristics and neighborhood characteristics are denoted by $\boldsymbol{X}_{i,t}$ and $\boldsymbol{Y}_{n(i),t}$ respectively and $E(\epsilon_{i,t+1} \mid \boldsymbol{X}_{i,t}, \boldsymbol{Y}_{n(i),t}) = 0$. Acceptance of the null hypothesis that $d' = \boldsymbol{0}$ is interpreted as acceptance of the null that no interaction effects exist. Examples of this type of regression include Brooks-Gunn et al. (1993), Corcoran et al. (1992), Rivkin (1997) and Zax and Rees (1998). These studies typically find some combinations of $\boldsymbol{Y}_{n(i),t}$ which are statistically significant, although there seems to be no consensus on which of these contextual effects are most robust. A useful extension of this work would be an analysis which explicitly attempted to identify robust neighborhood and individual controls, using techniques such as Leamer's (1983) extreme bounds analysis or Bayesian model averaging of the type advocated by Raftery (1995) and Raftery et al. (1997). While these procedures do not give a definitive solution to the problem of model uncertainty, they are nevertheless invaluable in clarifying dimensions along which arbitrary model assumptions (in this case choice of control variables) matters.

Several complementary strands exist to this class of empirical research. In one approach, the importance of neighborhood-level interactions effects on inequality is evaluated by assessing the effects on inequality measures of different sorting rules. This idea is originally due to Kremer (1997); nonlinear alternatives to Kremer's original analysis have been explored by Ioannides (1997b). In a second approach, the notion of neighborhoods has shifted from geographic proximity to membership in an ethnic group. Evidence of ethnic group effects has been found by Borjas (1992, 1995) and Bertrand et al. (1998). Similarly, Cutler and Glaeser (1997) illustrate how segregation adversely affects a number of socioeconomic outcomes for African Americans. A third strategy has been employed by Ioannides (1999) who shows how house spatial relations in price dynamics implicitly reveal neighborhood effects. In yet a fourth approach, Solon et al. (1999) use within-neighborhood correlations to overcome measurement problems associated with what neighborhood attributes actually matter for interactions. They find that once various family background variables are controlled for, within-neighborhood correlations in educational attainment are low.

Finally, there is a distinct literature on the relationship between interactions and efficient and/or equilibrium sorting. Becker (1973) and Sattinger (1975) are standard references; see Legros and Newman (1997) for the state-of-the-art. In addition, equilibrium sorting has been studied in many contexts using Tiebout type arguments. Recent contributions include Epple and Romer (1991) and Fernandez and Rogerson (1996) whose models are directly germane to the study of inequality. In an important paper, Epple and Sieg (1999) show how to econometrically implement models of this type. Our belief is that these types of models should be further employed to provide complementary insights to the main body of literature on interactions, as the strength of interaction effects should presumably be at least partially revealed by the neighborhood choices of individuals.

7.2. Analyses which are robust to unobserved correlated heterogeneity

From the perspective of empirical analysis, the main issue which has concerned researchers is the problem of spurious identification of interaction effects due to the likelihood of correlated unobservables existing among individuals in endogenously determined groups.

7.2.1. Matching

One approach to dealing with the possible unobserved correlated heterogeneity has attempted to identify environments which allow one to match populations subjected to different influences in order to assess the effect of changes in group membership. The most prominent type of matching study falls under the rubric of "natural experiments". By natural experiments, we refer to cases where interaction effects are identified by studying cases where some individuals that would normally be members of one group are moved to another through an exogenous intervention of some type. Those who are moved may be thought of as receiving a treatment, whereas those who remain may be thought of as a control group. While intuitively appealing, there are in fact many subtleties in analyzing data of this type. Heckman and Smith (1995), Heckman (1996, 1997), Heckman et al. (1998a,b), provide a wide ranging analysis of the salient issues. Hence an important future exercise is the reconsideration of some of these empirical studies in light of these recent econometric developments.

Among the most prominent examples of natural experiments of this type, we would list:

7.2.1.1. Gautreaux Assisted Housing Program. In 1966, the Chicago Housing Authority was sued for discrimination by public housing residents on the grounds that both the location of public housing sites and the allocation of slots in these sites intentionally placed minorities in isolated inner city neighborhoods. In an agreement worked out between the plaintiffs and defendants, known as the Gautreaux Assisted Housing Program, housing subsidies and placement services were established for public housing residents throughout Chicago. Rosenbaum (1995) and Yinger (1995) provide reviews of the details of the Gautreaux program. For the purposes of studying interactions, several points of these features are important. The number of participants each year was fixed and so, due to oversubscription, actual participants, after some screening, were randomly selected. Families who applied for assistance were randomly given a single option of moving to another part of Chicago or to moving to a suburb. (Families who declined the offered option were placed back in the pool of eligible families from which recipients of aid were drawn.)

A series of papers [Rosenbaum and Popkin (1991), Popkin et al. (1993), Rosenbaum (1995)], has analyzed the results of surveys of Gautreaux program participants in order to identify the effects of the differences between the urban and suburban environments on various socioeconomic measures.

While they are an important source of information on interactions effects, it is important to recognize that the Gautreaux data are not ideal for this purpose. Applicants to the program were dropped who either had poor rent paying histories or who failed a home inspection to determine whether they had mistreated their public housing. This prescreening eliminated approximately 30% of the program's applicants [Rosenbaum (1995)]. Further, the survey efforts conducted by Rosenbaum and coauthors exhibit some sample selection problems. In particular, those families who moved to suburbs and then returned to Chicago could not be identified. Hence, the evidence of neighborhood effects obtained from Gautreaux is, while informative, not decisive. That being said, recent work such as Rosenbaum et al. (1999), by linking

Gautreaux interview data to administrative data, should be able to partially address these concerns.

The Gautreaux program also illustrates the difficulty of identifying policy effects as well as a limitation in the utility of the naive estimator in predicting the effects of changes in interaction groups. Suppose that Gautreaux families are described by a linear-in-means model of the type:

$$\omega_i = d' \boldsymbol{Y}_{n(i)} + J m^e_{n(i)} + \epsilon_i. \tag{152}$$

Suppose that one new family is moved from the inner city to a suburb. In this case, the family's presence in the new location will have no effect on either $\boldsymbol{Y}_{n(i)}$ or $m^e_{n(i)}$ (and equivalently $m_{n(i)}$) in the new location and there will be no other Gautreaux families to reference on. In this case the knowledge of the parameters d' and J (which can be consistently estimated) and the neighborhood variables $\boldsymbol{Y}_{n(i)}$ and $m^e_{n(i)}$ in the old and new locations of residence will be sufficient to predict the effect on the family of the move.

However, suppose that the Gautreaux program is expanded to the extent that clusters of families are moved from an inner city to the new neighborhood. In this case, the appropriate model is

$$\omega_i = d' \boldsymbol{Y}_{n(i)} + J m^e_{n(i)} + d'_{G(i)} \boldsymbol{Y}_{n(i),\,G(i)} + J_{G(i)} m^e_{n(i),\,G(i)} + \epsilon_i. \tag{153}$$

Here, $G(i)$ denotes Gautreaux families in the neighborhood, so that for example, $m^e_{n(i),\,G(i)}$ denotes the mean behavior of Gautreaux families in a community. Predictions of the effect of a move of a cluster of families must therefore incorporate both the effects of the move on the mean for the neighborhood as a whole as well as the possibility that the Gautreaux families will represent a subgroup within the neighborhood which induces separate interactions. This means that the move of a cluster may be subject to social multipliers of the type we have described. At a minimum, the naive estimator is no longer useful for policy and prediction analysis. The analog of the self-consistency equation (21) must now be estimated along with the individual level equation (153) in order to permit predictions of the outcomes of cluster moves.

7.2.1.2. Moving to Opportunity Demonstration. The Moving to Opportunity Demonstration is an ongoing experimental demonstration being conducted by the Department of Housing and Urban Development to evaluate the effects of moving low-income families out of high-poverty neighborhoods; a detailed discussion of the program appears in Goering (1996). The demonstration randomly assigned a set of low income families normally eligible for Section 8 housing assistance vouchers to one of three groups: 1) those eligible for housing vouchers which are only usable in census tracts with less than 10% poverty, 2) those eligible for regular Section 8 vouchers with no locational restrictions, and 3) a group whose assistance is only based on residence in a public housing project. The demonstration is being conducted in 5 metropolitan areas: Baltimore, Boston, Chicago, Los Angeles and New York City. One motivation of the demonstration was a desire to address some of the self-selection problems associated with data from the Gautreaux Program. That being said, it is unclear at this stage to what extent self-selection is better controlled for here than in Gautreaux, given the voluntary nature of participation in the MTO demonstrations.

Preliminary results on the various experiments are becoming available. Ladd and Ludwig (1998) report evidence that those families in Baltimore that moved out of low income census tracts achieved access to superior schools as measured by a range of criteria. However, they find little evidence that the value added of these schools for the children in these families is higher than the schools used by families in the comparison and control groups. For the Boston demonstration, Katz et al. (1997) also find evidence that the MTO program has been successful in generating relocation of families, this time defined as movements out of low poverty neighborhoods. They also find that children in both types of families eligible for vouchers exhibited substantially higher test scores as well as lower incidences of behavioral problems.

7.2.1.3. Milwaukee School Voucher Program. In 1990, Wisconsin implemented the nation's first public school voucher program. In essence, this program made available school vouchers equal to the average per pupil expenditure by public schools in Milwaukee. Applicants to the program were required to fulfill several criteria. Oversubscription to the program has meant that a random subset of eligible applicants have actually been able to participate in the program. Eligibility for the program was restricted by two criteria: 1) a family could not have an annual income which exceeded 1.75 times the poverty line, and 2) the student to receive the voucher could not have previously been enrolled in a private school in the year prior to the use of the voucher.

As the number of applicants greatly exceeded the number of available vouchers, the randomness of the selection process meant that there existed two groups of students, namely those who did or did not receive vouchers for private schools, whose subsequent performance could be compared. Rouse (1998a,b) and Witte (1997) have both studied this question. Interestingly, they have come to quite different conclusions concerning the effects of private versus public schools on education. Rouse concludes that there are some benefits to private schools in this sample whereas Witte does not. These differences appear to stem from different choices concerning the appropriate control group for analysis. Rouse uses those students who applied but were not selected for the program whereas Witte uses citywide average student outcomes. In terms of differencing out characteristics of the control and treatment groups, Rouse's approach seems clearly correct.

While important with respect to the issue of school vouchers and public policy, there are several grounds for supposing that the Milwaukee evidence has limited implications in terms of adducing the importance of interactions. As both Rouse and Witte are aware, since the majority of participants in the program went to one of only three different schools, the generality of any of the results is questionable.

Further, it is important to remember that interactions, as conventionally understood, may not explain the differences either here or for differences between public and Catholic schools, which are discussed below. Differences in disciplinary standards or teacher expectations could potentially explain differences with the public schools independent of any interactions effects such as peer group influences. An interesting question for future research is therefore the determination of whether observed school differences occur due to interactions between students or due to alternative educational and disciplinary standards.

7.2.1.4. Classroom tracking. A standard problem in school organization is whether students should be tracked, i.e., segregated by ability and/or achievement across classes. A number of classroom experiments have been conducted in which educational outcomes for students tracked by initial measures of ability of achievement are compared to students who are randomly assigned to classrooms.

One such experiment occurred in Montreal and has been analyzed by Henderson et al. (1978). This paper analyzes data from French speaking students in Montreal in which children who were segregated on the basis of IQ tests administered in kindergarten and students who were randomly assigned to classes were compared in terms of achievement in grades 1–3. Henderson, Mieszkowski and Sauvageau found significant effects from this type of classroom tracking. Interestingly, while randomization raised overall average performance, there was a clear diminution of the performance of students with higher test scores under random assignments. Hence randomization involves both redistribution as well as an increase in average achievement scores.

Unsurprisingly, ability grouping has also been studied quite extensively by education researchers, and has been a source of considerable controversy within the education literature. Slavin (1990) reviews a large number of tracking versus random assignment experiments in high schools and concludes that for secondary students "... between-class ability grouping plans have little or no effect on ... achievement ... at least as measured by standardized tests" (p. 494). However, even this survey conclusion has been disputed by other education scholars as evidenced in the commentaries on that article. Our limited survey of the education literature suggests that there is little decisive evidence on this question, and that many of the studies are plagued by poor controls for individual characteristics; further, much of this literature seems laden with political concerns on the parts of researchers which make the assessment of the statistical analysis problematic. These techniques may also facilitate the determination of which neighborhood characteristics are relevant in generating interaction effects. Weinberg et al. (1999) show that an analysis employing a broad range of possible neighborhood controls can lead one to reject peer group and role model effects in favor of broader socioeconomic characteristics as the determinant of neighborhood effects.

7.2.1.5. Siblings. Matching comparisons have also been employed to directly control for unobserved family effects. Aaronson (1997, 1998) proposes the use of sibling data to difference out unobserved family characteristics. This is possible under the assumption that the unobservable characteristics are constant within a family across time. He then identifies sibling pairs from the National Longitudinal Survey of Youth in which one sibling was exposed to a different neighborhood than another. This allows him to estimate models of differences in sibling outcomes which include differences in neighborhood characteristics. This estimation strategy is therefore equivalent to the standard one in panel data studies of differencing out unobserved fixed effects. Plotnick and Hoffman (1996) apply the same idea to a sample of sisters from the Panel Study of Income Dynamics and consider both continuous and discrete outcomes. Exploiting Chamberlain (1984) in order to eliminate unobserved fixed effects for binary choices, they find little evidence of neighborhood effects with respect to either out of wedlock births or any post-secondary education. This study finds no evidence of neighborhood effects on a particular income measure.

7.2.2. Instrumental variables

Rather than employ data sets where interaction effects can be identified through the comparison of otherwise equivalent treatment and control groups, there has been a parallel literature which has tried to use more conventional econometric methods to deal with unobserved correlates.

7.2.2.1. Neighborhood socioeconomic influences. Evans et al. (1992) appears to be the first study of neighborhood influences which formally accounts for the endogeneity of neighborhood residence. The analysis is specifically concerned with identifying the role of neighborhood characteristics on the probability of teen pregnancy. Using a probit framework, this probability is assumed to depend on both a range of individual characteristics as well as a variable which is the logarithm of the percentage of other students in an individual's high school who are categorized as "disadvantaged" as defined under guidelines of the Elementary and Secondary Education Act. In probit regressions which treat this measure as exogenous, this measure of disadvantaged schoolmates is shown to statistically significantly increase the probability of a teen pregnancy.

In order to deal with the possibility that the neighborhood characteristic measure is correlated with an unobserved individual characteristic, as would occur if parental quality is negatively associated with the neighborhood characteristic, Evans, Oates, and Schwab propose four instrumental variables each of which is measured at the level of the metropolitan area in which the secondary student lives: 1) the unemployment rate, 2) median family income, 3) the poverty rate, and 4) the percentage of adults who are college graduates. The implicit assumption in this analysis is that the metropolitan area of residence is exogenous for families, although location within a metropolitan area is a choice variable. Employing these instruments, the contextual effect found in the univariate analysis disappears both in terms of magnitude and in terms of statistical significance.

7.2.2.2. Catholic versus public schools. Starting with Coleman et al. (1982), a number of authors have studied the reasons why student performance in Catholic schools is on average superior to that found in public equivalents. A critical issue in evaluating the implications of this fact is determining whether the differences are due to self-selection with respect to school enrollment versus something about differences in the school environment per se.

Evans and Schwab (1995) and Neal (1997) attempt to deal with the effect of self-selection by identifying instrumental variables which correlate with Catholic school choice but do not correlate with unobservable individual characteristics which would lead to better school performance. Neal's analysis seems especially comprehensive. He proposes two instruments which plausibly correlate with the decision to attend a Catholic school but not with unobserved individual characteristics which would lead to superior academic performance regardless of which school was attended: 1) the fraction of Catholic in county of residence population, which should correlate with tuition costs since higher percentages lead to greater Church subsidies to schools, and 2) the number of Catholic secondary schools per square mile within county, which should correlate negatively with transportation costs. The idea is tuition and transportation costs are plausibly correlated with the determinants of Catholic school choice without being correlated with an unobserved student quality variable. Neal finds that there are substantial educational gains for urban minorities who attend Catholic schools, but not for suburban students or whites in general.

8. Summary and conclusions

This chapter illustrates both the progress which has been made in utilizing interactions to understand economic phenomena as well as the many areas in which further research is required. In our judgment, there currently exists a good understanding of static interactions-based models both in terms of theory and econometrics. However, the empirical literature, while containing many insightful approaches to uncovering interactions, has yet to exploit a full structural estimation approach. Such a step is particularly important if one wishes to identify the presence of multiple equilibria. Further, there exist a number of areas in terms of theory and econometric methodology which have yet to be fully examined. Three examples come readily to mind. First, the analysis of dynamic interaction models with endogenous neighborhood formation and their panel data analogs is still in its infancy. This analysis, fortunately, should have useful antecedents in the urban economics literature such as Miyao (1978). Second, the theoretical models of interactions currently treat the sources of interactions as a black box. In understanding phenomena such as social norms or culture, this is clearly inadequate; see the interesting analysis of Emirbayer and Goodwin (1997) for a discussion of the importance of properly accounting for the microfoundations of norms and culture. Third, the econometric literature has almost exclusively concentrated on global interactions, and so the analysis of identification and estimation needs to be extended to alternative interaction structures. Therefore, we are very confident that interactions-based models will continue to prove to be a productive area of research for methodologists and empiricists alike.

Appendix A

A.1. Properties of binary choices made under a social planner

A.1.1. Basics

Recalling the discussion in section 2.5 in the text we consider a population of I individuals whose choices, as determined by a social planner, follow the probability model

$$\mu(\boldsymbol{\omega}) = \exp\Big(\beta\Big(\sum_i \left(u(\omega_i, \boldsymbol{Z}_i) + S(\omega_i, \boldsymbol{Z}_i, \boldsymbol{\omega}_{-i})\right)\Big)\Big)/Z_I. \tag{A.1}$$

In this case, $\boldsymbol{\omega}_{-i}$ is substituted for $\mu_i^e(\boldsymbol{\omega}_{-i})$ in the social utility terms of the noncooperative problems and Z_I is a normalizing constant. For the case of symmetric global interactions ($J_{i,j} = \frac{J}{I}$), employing the same transformations as done in the noncooperative case means that this probability may be rewritten as

$$\mu(\boldsymbol{\omega}) = \exp\Big(\beta\Big(\sum_i h_i\omega_i + \frac{J}{2I}\Big(\sum_i \omega_i\Big)^2\Big)\Big)/Z_I, \tag{A.2}$$

where the normalizing constant Z_I is

$$\sum_{\nu_1 \in \{-1,1\}} \cdots \sum_{\nu_I \in \{-1,1\}} \exp\Big(\beta\Big(\sum_{i=1}^{I} h_i\nu_i + \frac{J}{2I}\Big(\sum_{i=1}^{I} \nu_i\Big)^2\Big)\Big). \tag{A.3}$$

This equation corresponds to Equation (45) in the text.

In order to analyze this model, we make use of the following identity

$$\exp\left(a^2\right) = (2\pi)^{-1/2} \int_{-\infty}^{\infty} \exp\left(-\frac{x^2}{2} + \sqrt{2}xa\right) \mathrm{d}x. \tag{A.4}$$

This identity can be verified immediately by dividing both sides of the expression by $\exp(a^2)$ and recalling that the integral of the probability density of a normal $(\sqrt{2}a, 1)$

random variable over its support is 1. Using the change of variable $y = x\left(\frac{\beta J}{I}\right)^{1/2}$, it must be the case that

$$\mu(\boldsymbol{\omega}) = \left(\frac{I}{2\pi\beta J}\right)^{1/2} \int_{-\infty}^{\infty} \exp\left(-\frac{y^2 I}{2\beta J}\right) \prod_i \exp\left(\left(y + \beta h_i\right)\omega_i\right) \mathrm{d}y/Z_I, \tag{A.5}$$

where

$$Z_I = \left(\frac{I}{2\pi\beta J}\right)^{1/2} \int_{-\infty}^{\infty} \exp\left(-\frac{y^2 I}{2\beta J}\right) \prod_i M\ \left(y + \beta h_i\right) \mathrm{d}y, \tag{A.6}$$

and

$$M(s) = \exp(s) + \exp(-s). \tag{A.7}$$

Notice that

$$\int_{-\infty}^{\infty} \exp\left(-\frac{y^2 I}{2\beta J}\right) \prod_i M\left(y + \beta h_i\right) \mathrm{d}y = \int_{-\infty}^{\infty} \exp\left(I H_I(y)\right) \mathrm{d}y, \tag{A.8}$$

where

$$H_I(y) = \frac{1}{I} \sum_i \ln\left(\exp\left(\upsilon_i(1)\right) + \exp\left(\upsilon_i(-1)\right)\right), \tag{A.9}$$

and

$$\upsilon_i(\omega_i) = \beta h_i \omega_i + y\omega_i - \frac{y^2}{2\beta J}. \tag{A.10}$$

Notice that if $h_i = h\ \forall\ i$, then $H_I(y)$ does not depend on I.

It is shown rigorously by Amaro de Matos and Perez (1991) that as $I \Rightarrow \infty$, integrals of the form

$$\int_{-\infty}^{\infty} \exp\left(I H_I(y)\right) \mathrm{d}y, \tag{A.11}$$

"pack" all mass onto the global maximizing point

$$y^* = \arg\max_y\left(H(y)\right), \tag{A.12}$$

where

$$H(y) = \int_{-\infty}^{\infty} \ln\left(\exp\left(\beta h + y - \frac{y^2}{2\beta J}\right) + \exp\left(-\beta h - y - \frac{y^2}{2\beta J}\right)\right) \mathrm{d}F(h), \tag{A.13}$$

and $F(h)$ is the cumulative distribution function of h. One therefore expects (and can prove) that

$$y_I^* = \arg\max_y (H_I(y)) \Rightarrow y^* = \arg\max_y (H(y)). \tag{A.14}$$

Simple algebra reveals that the first order condition for the maximum of $H(\beta Jm)$ over m is

$$m = \int_{-\infty}^{\infty} \tanh(\beta h + \beta Jm)\,\mathrm{d}F(h). \tag{A.15}$$

When $h_i = h \forall i$, this equation also holds for the expected value of each individual i and hence the sample average by symmetry, which gives us Theorem 4.

Finally, this result suggests that if we replace the integral over y in Equation (A.5) with a Dirac delta function whose mass is at y^*, we can obtain an approximate probability for the system of the form

$$\mu(\boldsymbol{\omega}) = \left(\frac{I}{2\pi\beta J}\right)^{1/2} \exp\left(-\frac{y^{*2}I}{2\beta J}\right) \prod_i \exp((y^* + \beta h_i)\,\omega_i)/Z_I, \tag{A.16}$$

where Z_I is a normalizing constant.

A.1.2. Asymptotic moments: proof of Theorem 11

Let $E(\omega_i)$ denote the expectation of ω_i with respect to the probability measure (A.2). In order to determine the behavior of sample averages as $I \Rightarrow \infty$, we again consider the case where $h_i = h \forall i$. Notice that the argument of the previous section implies that

$$\lim_{I\Rightarrow\infty} E(\omega_i) = \frac{\lim_{I\Rightarrow\infty} \int_{-\infty}^{\infty} \exp(IH_1(y))\,G_1(y)\,\mathrm{d}y}{\lim_{I\Rightarrow\infty} \int_{-\infty}^{\infty} \exp(IH_1(y))\,\mathrm{d}y,} \tag{A.17}$$

where

$$H_1(y) = \ln(M(\beta h + y)) - \frac{y^2}{2\beta J}, \tag{A.18}$$

and

$$G_1(y) = \frac{\exp(\beta h + y) - \exp(-\beta h - y)}{M(\beta h + y)} = \frac{M'(\beta h + y)}{M(\beta h + y)}. \tag{A.19}$$

We employ LaPlace's method [Kac (1968, p. 248) or Ellis (1985, pp. 38, 50–51)] to obtain the limiting values of these integrals. As described above, intuitively, all mass in these integrals gets packed onto the global maximizer y^*.

We restate the following useful result which is proven in Murray (1984, p. 34).

Approximation theorem. *Let $H(t)$ be a function on the interval (a, b) which takes a global maximum at a point α in the interval and let $H(t)$ be smooth enough to possess a second-order Taylor expansion at point α with $H''(\alpha) < 0$. Let $G(t)$ denote a continuous function. Then*

$$\int_{-\infty}^{\infty} G(t)\exp(IH(t))\,\mathrm{d}t = \exp(IH(\alpha))\,G(\alpha)\left(\frac{-2\pi}{IH''(\alpha)}\right)^{1/2} + O\left(I^{-3/2}\right). \tag{A.20}$$

This formula states, in a precise way, the sense in which the mass of the integral piles up at the maximizer α as $I \Rightarrow \infty$. Using this formula, letting $\alpha = y$,

$$\frac{\int_{-\infty}^{\infty}\exp(IH_1(y))\,G_1(y)\,\mathrm{d}y}{\int_{-\infty}^{\infty}\exp(IH_1(y))\,\mathrm{d}y} = \frac{\exp(IH_1(\alpha))\,G_1(\alpha)\left(\frac{-2\pi}{IH_1''(\alpha)}\right)^{1/2} + O\left(I^{-3/2}\right)}{\exp(IH_1(\alpha))\left(\frac{-2\pi}{IH_1''(\alpha)}\right)^{1/2} + O\left(I^{-3/2}\right)}, \tag{A.21}$$

which is easily seen to converge to $G_1(\alpha)$ as $I \Rightarrow \infty$. Hence we have

$$\begin{aligned}\lim_{I\Rightarrow\infty} E(\omega_i) = G_1(y^*) = m^* &= \frac{\exp(\beta h + y^*) - \exp(-\beta h - y^*)}{M(\beta h + y^*)}\\ &= \frac{M'(\beta h + y^*)}{M(\beta h + y^*)} = \tanh(\beta h + y^*) = \tanh(\beta h + \beta J m^*),\end{aligned} \tag{A.22}$$

where $y^* = \beta J m^*$.

The problem that m^* solves appears mysterious at first glance. However, there is an interesting connection between our solution to the behavior of a social planner and the maximization of social surplus as analyzed in McFadden (1981, Chapter 5). Following McFadden, social surplus will equal $\sum_i (u(\omega_i, \boldsymbol{X}_i) - \frac{J}{2}(\omega_i - \bar{\omega}_I)^2)$. If all agents have common characteristics $\boldsymbol{X}_i$, then following Equation (A.2), the probability of the social surplus can be expressed as a function of $G(\boldsymbol{\omega}) = \sum_i h\omega_i + \frac{J}{2I}\left(\sum_i \omega_i\right)^2$. Then it can be shown [Brock (1993)] that

$$\begin{aligned}&\beta\left(\lim_{I\Rightarrow\infty} E\left(\max_{\boldsymbol{\omega}} I^{-1} G(\boldsymbol{\omega})\right)\right)\\ &\quad= \lim_{I\Rightarrow\infty}\left(I^{-1}\ln(Z_I)\right) = \max_{y}\ln\left(\exp\left(-\frac{y^2}{2\beta J}\right)M(\beta h + y)\right)\\ &\quad= \max_{m}\ln\left(\exp\left(-\frac{(\beta J m)^2}{2\beta J}\right)M(\beta h + \beta J m)\right).\end{aligned} \tag{A.23}$$

As would be expected, one maximizes a notion of social welfare in the large economy limit in order to find the socially optimal states.

Now that the expected value for each choice has been analyzed, we can consider laws of large numbers for data generated in this environment. First, we consider the sample mean, $\bar{\omega}_I = I^{-1}\sum_i \omega_i$. Notice that the limiting behavior of the sample mean in distribution ($\Rightarrow_d$) can be inferred from weak convergence ($\Rightarrow_\omega$) since weak convergence necessarily implies convergence in distribution [see Lukacs (1975, p. 9) for a typical proof]. By Tchebychev's inequality,

$$\mu\left(|\bar{\omega}_I - m^*| \geqslant \epsilon\right) \leqslant \frac{\operatorname{Var}(\bar{\omega}_I - m^*)}{\epsilon^2}, \tag{A.24}$$

so it is sufficient to prove $\lim_{I\Rightarrow\infty} \operatorname{Var}(\bar{\omega}_I - m^*) = 0$. To do this, it is sufficient to show that $I^{-2}\sum_i \omega_i \sum_j \omega_j \Rightarrow_\omega m^{*2}$. However, this can be verified (after considerable algebra) by computing $I^{-2}\sum_i \omega_i \sum_j \omega_j$ directly and using LaPlace's method as employed in Murray above to verify that $I^{-2}\sum_i \omega_i \sum_j \omega_j \Rightarrow_\omega \tanh(\beta h + y^*)^2 = m^{*2}$. This proves Theorem 12.

A.1.3. Maximum likelihood theory

Consider $g = 1 \cdots G$ distinct neighborhoods with observations $\boldsymbol{X}_{i,g}, i = 1 \cdots I$ and $\bar{\omega}_g = I^{-1}\sum_i \omega_{i,g}$ available for each g. Define the likelihood function for the data from these neighborhoods as $\prod_g \mu\left(\boldsymbol{\omega}_g\right)$ where $\boldsymbol{\omega}_g = \left(\omega_{1,g} \cdots \omega_{I,g}\right)$. When choices are consistent with the solution to a social planners problem, the likelihood function within each neighborhood will have the form

$$\mu\left(\boldsymbol{\omega}_g\right) \sim \exp\Bigl(\sum_i \Bigl(\tfrac{1}{2}c'\boldsymbol{X}_{i,g} + \frac{J}{I}\Bigl(\sum_j \omega_{j,g}\Bigr)\Bigr)\omega_{i,g}\Bigr), \tag{A.25}$$

which can be rewritten as

$$\left(\frac{I}{2\pi\beta J}\right)^{1/2} \int_{-\infty}^{\infty} \exp\left(-\bar{\omega}_g^2 \frac{IJ}{2}\right) \prod_i \exp\left(\left(\tfrac{1}{2}c'\boldsymbol{X}_{i,g} + J\bar{\omega}_g\right)\omega_{i,g}\right) \mathrm{d}\bar{\omega}_g. \tag{A.26}$$

Define the parameter vector $\theta = (c, J)$. One can consider the mean log likelihood over all observations

$$\frac{1}{GI}\sum_g \ln\left(\mu\left(\boldsymbol{\omega}_g\right)\right). \tag{A.27}$$

For large I, and letting $F(x) = \frac{\exp(x)}{1+\exp(x)}$ the density of this likelihood will approximately equal

$$\begin{aligned}\frac{1}{G}\sum_g \frac{1}{I}\sum_i \Bigl[&\left(\frac{1+\omega_{i,g}}{2}\right) \ln\left(F\left(c'\boldsymbol{X}_{i,g} + 2J\mu_{I,g}\right)\right) \\ &+ \left(\frac{1-\omega_{i,g}}{2}\right) \ln\left(F\left(-\left(c'\boldsymbol{X}_{i,g} + 2J\mu_{I,g}\right)\right)\right)\Bigr],\end{aligned} \tag{A.28}$$

where

$$\mu_{I,g} = \arg\max\left(H_{I,g}(\mu)\right), \tag{A.29}$$

and

$$H_{I,g}(\mu) = -\frac{\mu^2 J}{2} + \frac{1}{I}\left(\sum_i \ln\left(\exp\left(\tfrac{1}{2}c'\boldsymbol{X}_{i,g} + J\mu\right)\right) + \exp\left(-\tfrac{1}{2}c'\boldsymbol{X}_{i,g} - J\mu\right)\right) \tag{A.30}$$

Note that $H_{I,g}$ converges to

$$H_g(\mu) = -\frac{\mu^2 J}{2} + \int \ln\left(\exp\left(\tfrac{1}{2}c'\boldsymbol{X}_{i,g} + J\mu\right) + \exp\left(-\tfrac{1}{2}c'\boldsymbol{X}_{i,g} - J\mu\right)\right) \mathrm{d}F_g\left(\boldsymbol{X}_{i,g}\right), \tag{A.31}$$

so that under regularity conditions such as those described in Newey and McFadden (1994, p. 2121) it must be the case that

$$\mu_{I,g} \Rightarrow_w \mu_g = \arg\max\left(H_G(\mu)\right). \tag{A.32}$$

Notice that the naive estimator introduced in Section 3 inserts $\bar{\omega}_g$ in place of $\mu_{I,g}$ in the sample likelihood (A.28) above and selects θ to maximize the modified sample log likelihood function. This means that the naive estimator does not allow the data to directly address the possibility of discontinuous neighborhood responses because the standard maximum likelihood theory in logistic models yields a strictly concave optimization problem. Hence the optimization problem will be continuous in parameters such as the distribution function of individual characteristics, $F_g\left(\boldsymbol{X}_{i,g}\right)$.

This suggests that one might wish to modify this log likelihood by adding a penalty function of the form

$$\frac{A}{G}\sum_g \left(\bar{\omega}_g - \mu_{I,g}\right)^2. \tag{A.33}$$

$A = 0$ will correspond to the naive estimator. Intuitively, as A increases the penalty will push the parameter estimates towards those of the complete estimator, i.e., one which accounts for the relationship between the neighborhood characteristics and neighborhood mean behavior.

A.2. Proof of Theorem 5

For a given parameter set (k, c, d, J), assume by way of contradiction that there exists an alternative $\left(\bar{k}, \bar{c}, \bar{d}, \bar{J}\right)$ such that on supp$(\boldsymbol{X}, \boldsymbol{Y}, m^e)$ we have

$$\left(k - \bar{k}\right) + \left(c' - \bar{c}'\right)\boldsymbol{X}_i + \left(d' - \bar{d}'\right)\boldsymbol{Y}_{n(i)} + \left(J - \bar{J}\right) m^e_{n(i)} = 0, \tag{A.34}$$

and

$$\begin{aligned} m^e_{n(i)} = m_{n(i)} &= \int \omega_i \mathrm{d}F\left(\omega_i \mid k + c'\boldsymbol{X} + d'\boldsymbol{Y}_{n(i)} + Jm_{n(i)}\right) \mathrm{d}F_{\boldsymbol{X} \mid \boldsymbol{Y}_{n(i)}} \\ &= \int \omega_i \mathrm{d}F\left(\omega_i \mid \bar{k} + \bar{c}'\boldsymbol{X} + \bar{d}'\boldsymbol{Y}_{n(i)} + \bar{J}m_{n(i)}\right) \mathrm{d}F_{\boldsymbol{X} \mid \boldsymbol{Y}_{n(i)}}. \end{aligned} \tag{A.35}$$

Notice the Proposition is true if it is the case that $J - \bar{J}$ is zero. Otherwise $\boldsymbol{X}_i$, and $\boldsymbol{Y}_{n(i)}$ would lie in a proper linear subspace of R^{r+s} which violates Assumption *i*. Equation (A.34) implies that for elements of supp($\boldsymbol{X}$, $\boldsymbol{Y}$, m^e), conditional on $\boldsymbol{Y}_{n(i)}$

$$\left(c' - \bar{c}'\right) \boldsymbol{X}_i = \rho\left(\boldsymbol{Y}_{n(i)}\right), \tag{A.36}$$

where $\rho(\boldsymbol{Y}_{n(i)}) = -\left(k - \bar{k}\right) - \left(d' - \bar{d}'\right) \boldsymbol{Y}_{n(i)} - \left(J - \bar{J}\right) m^e_{n(i)}$. Equation (A.36) must hold for all neighborhoods, including n_0 as described in Assumption *iv* of the Theorem. This would mean that, conditional on $\boldsymbol{Y}_{n_0}$, and given that $\boldsymbol{X}_i$ cannot contain a constant by Assumption *iii*, that $\boldsymbol{X}_i$ is contained in a proper linear subspace of R^r and therefore violates Assumption *iv* of the Proposition. Hence, c is identified.

Given identification of c, Equation (A.34) now implies, if $J \neq \bar{J}$, that $m^e_{n(i)}$ is a linear function of $\boldsymbol{Y}_{n(i)}$, unless $\left(d' - \bar{d}'\right)$ and/or $m^e_{n(i)}$ is always equal to zero. The latter is ruled out by Assumption *vi*. Linear dependence of $m^e_{n(i)}$ on $\boldsymbol{Y}_{n(i)}$ when $\left(d' - \bar{d}'\right) \neq 0$ contradicts the combination of the requirement that support of $m^e_{n(i)}$ is $[-1, 1]$ with Assumption *v*, that the support of each component of $\boldsymbol{Y}_{n(i)}$ is unbounded, since $\boldsymbol{Y}_{n(i)}$ can, if it is unbounded, assume values with positive probability that violate the bounds on $m^e_{n(i)}$. So, J is identified. If J is identified and $\left(d' - \bar{d}'\right) \neq 0$, then Equation (A.34) requires that

$$\left(d' - \bar{d}'\right) \boldsymbol{Y}_{n(i)} = -\left(k - \bar{k}\right), \tag{A.37}$$

for all $\boldsymbol{Y}_{n(i)} \in \operatorname{supp}\left(\boldsymbol{Y}_{n(i)}\right)$. This implies, since by Assumption *iii* $\boldsymbol{Y}_{n(i)}$ does not contain a constant, that $\operatorname{supp}\left(\boldsymbol{Y}_{n(i)}\right)$ is contained in a proper linear subspace of R^s, which contradicts condition *ii* of the Theorem. Therefore, $d' = \bar{d}'$. This immediately implies that $k = \bar{k}$ and the Theorem is verified.

A.3. Proof of Theorem 7

As is done in the text, A denotes the parameter set (k, c, d, J) and the conditional mean function is $H = k + c'\boldsymbol{X}_i + d'\bar{\boldsymbol{X}}_{n(i)} + JG(m)$. To verify the theorem, it is necessary to show that the components of the gradient vector

$$\mathrm{d}_A H = \frac{\partial H}{\partial A} + \frac{\partial H}{\partial m}\frac{\partial m}{\partial A}, \tag{A.38}$$

define a linearly independent collection of functions of $\boldsymbol{X}_i$ and $\bar{\boldsymbol{X}}_{n(i)}$ on $\operatorname{supp}\left(\boldsymbol{X}_i, \bar{\boldsymbol{X}}_{n(i)}\right)$. Differentiation implies the following, which we will use,

$$\frac{\partial H}{\partial A} = \left(1, \boldsymbol{X}_i, \bar{\boldsymbol{X}}_{n(i)}, G(m)\right), \tag{A.39}$$

$$\frac{\partial H}{\partial m} = J\left(1 + \xi \frac{\mathrm{d}g(m)}{m}\right). \tag{A.40}$$

Since $J \neq 1$ and g is C^2, the neighborhood N_ϵ can always be chosen so that an implicit function $m\left(\bar{\boldsymbol{X}}_{n(i)}, A, \xi\right)$ exists. Also, define the function $J(m, \xi) = J\left(1 + \xi \frac{\mathrm{d}g(m)}{m}\right)$.

Rewrite the gradient as

$$\mathrm{d}_A H = \frac{1}{1 - J(m, \xi)}\left(1, \boldsymbol{X}_i + J(m, \xi)\left(\bar{\boldsymbol{X}}_{n(i)} - \boldsymbol{X}_i\right), \bar{\boldsymbol{X}}_{n(i)}, m + \xi g(m)\right). \tag{A.41}$$

If ξ is close enough to zero, $J(m, \xi)$ cannot equal 1 since $J \neq 1$ by Assumption *iii*. This is a vector proportional to the form $\upsilon = (1, \upsilon_2\left(\boldsymbol{X}_i, \bar{\boldsymbol{X}}_{n(i)}\right), \upsilon_3\left(\bar{\boldsymbol{X}}_{n(i)}\right), \upsilon_4\left(\bar{\boldsymbol{X}}_{n(i)}\right))$. Notice that we have eliminated m since its implicit function solution makes it a function of $\bar{\boldsymbol{X}}_{n(i)}$. In order to show linear independence, we must verify that

$$a_1 + a_2\upsilon_2\left(\boldsymbol{X}_i, \bar{\boldsymbol{X}}_{n(i)}\right) + a_3\upsilon_3\left(\bar{\boldsymbol{X}}_{n(i)}\right) + a_4\upsilon_4\left(\bar{\boldsymbol{X}}_{n(i)}\right) = 0, \tag{A.42}$$

implies that $a_1 = a_2 = a_3 = a_4 = 0$.

Since only υ_2 depends on $\boldsymbol{X}_i$, Equation (A.42) can only hold if $a_2 = 0$; otherwise Assumption *ii* would be violated. Further, if $a_4 = 0$, then Assumption *i* is violated. This is true because $\upsilon_3\left(\bar{\boldsymbol{X}}_{n(i)}\right)$ is proportional to $\bar{\boldsymbol{X}}_{n(i)}$. We can therefore, without loss of generality assume $a_4 = -1$.

The condition for linear independence can now be written as

$$m\left(\bar{\boldsymbol{X}}_{n(i)}, A, \xi\right) + \xi g\left(m\left(\bar{\boldsymbol{X}}_{n(i)}, A, \xi\right)\right) = a_1 + a_3\bar{\boldsymbol{X}}_{n(i)}. \tag{A.43}$$

We pair this with the self-consistency condition written as

$$m\left(\bar{\boldsymbol{X}}_{n(i)}, A, \xi\right) = k + \left(c' + d'\right)\bar{\boldsymbol{X}}_{n(i)} + J\left(m\left(\bar{\boldsymbol{X}}_{n(i)}, A, \xi\right)\right) + \xi g\left(m\left(\bar{\boldsymbol{X}}_{n(i)}, A, \xi\right)\right). \tag{A.44}$$

We will verify that Equations (A.43) and (A.44) lead to a contradiction when $\frac{\mathrm{d}g}{\mathrm{d}m}$ differs across any two m values, say m_1 and m_2. Since at least two such values must exist by Assumption *iv*, this will complete the proof.

On the open set O described by Assumption *iv*, we can differentiate both these equations with respect to $\bar{\boldsymbol{X}}_{n(i)}$, obtaining

$$\left(1+\xi\frac{\mathrm{d}g\left(m\left(\bar{\boldsymbol{X}}_{n(i)},A,\xi\right)\right)}{\mathrm{d}m}\right)\frac{\mathrm{d}m\left(\bar{\boldsymbol{X}}_{n(i)},A,\xi\right)}{\mathrm{d}\bar{\boldsymbol{X}}_{n(i)}}=a_3, \tag{A.45}$$

and

$$\frac{\mathrm{d}m\left(\bar{\boldsymbol{X}}_{n(i)},A,\xi\right)}{\mathrm{d}\bar{\boldsymbol{X}}_{n(i)}}\left(1-J\left(1+\xi\frac{\mathrm{d}g\left(m\left(\bar{\boldsymbol{X}}_{n(i)},A,\xi\right)\right)}{\mathrm{d}m}\right)\right)=\left(c'+d'\right). \tag{A.46}$$

Equating $\frac{\mathrm{d}m\left(\bar{\boldsymbol{X}}_{n(i)},A,\xi\right)}{\mathrm{d}\bar{\boldsymbol{X}}_{n(i)}}$ across these expressions yields

$$\left(1-J\left(1+\xi\frac{\mathrm{d}g\left(m\left(\bar{\boldsymbol{X}}_{n(i)},A,\xi\right)\right)}{\mathrm{d}m}\right)\right)a_3-(c+d)\left(1+\xi\frac{\mathrm{d}g\left(m\left(\bar{\boldsymbol{X}}_{n(i)},A,\xi\right)\right)}{\mathrm{d}m}\right)=0, \tag{A.47}$$

or

$$\xi\frac{\mathrm{d}g\left(m\left(\bar{\boldsymbol{X}}_{n(i)},A,\xi\right)\right)}{\mathrm{d}m}\left((c+d)+Ja_3\right)+\left((c+d)+Ja_3-a_3\right)=0. \tag{A.48}$$

Recall that ξ and $\frac{\mathrm{d}g}{\mathrm{d}m}$ are scalars, whereas c, d, and a_3 are $r \times 1$ vectors. By construction of g, we have the existence of two values of m, call them m_1 and m_2, such that in the population data, $\frac{\mathrm{d}g}{\mathrm{d}m}$ differs across them. Applying this component by component to Equation (A.47), one can show that this implies that $(c+d)+Ja_3 = 0$. By Equation (A.48), this means that $a_3 = 0$. But from Equation (A.49), this would imply that

$$m\left(\bar{\boldsymbol{X}}_{n(i)},A,\xi\right)+\xi g\left(m\left(\bar{\boldsymbol{X}}_{n(i)},A,\xi\right)\right)=a_1. \tag{A.49}$$

But this would contradict the part of Assumption *iv* that $m_{n(i)}$ in the data is nonconstant. Therefore, the model and assumptions described by the Theorem require that the components of the gradient (A.38) are linearly independent when $\xi \neq 0$. Notice that when $\xi = 0$, the gradient will not be of full rank, because $m\left(\bar{\boldsymbol{X}}_{n(i)},A,0\right)$ is linear in $\bar{\boldsymbol{X}}_{n(i)}$. Hence the local nonidentification of the linear-in-means model can be perturbed away by a C^2-small change from Jm to $Jm+\xi g(m)$, which completes the proof.

References

Aaronson, D. (1997), "Sibling estimates of neighborhood effects", in: J. Brooks-Dunn, G. Duncan and L. Aber, eds., Neighborhood Poverty: Policy Implications for Studying Neighborhoods, Vol. II (Russell Sage Foundation, New York).

Aaronson, D. (1998), “Using sibling data to estimate the impact of neighborhoods on children’s educational outcomes”, Journal of Human Resources XXXIII:915–946.

Ahn, H., and J.L. Powell (1993), “Semiparametric estimation of censored regression models with a nonparametric selection mechanism”, Journal of Econometrics 58:3–29.

Akerlof, G. (1997), “Social distance and social decisions”, Econometrica 65:1005–1028.

Akerlof, G., and R. Kranton (1998), “Identity and economics”, Mimeo (University of California at Berkeley).

Akerlof, G., and J. Yellen (1994), “Gang behavior, law enforcement, and community values”, in: H. Aaron, T. Mann and T. Taylor, eds., Values and Public Policy (Brookings Institution Press, Washington, DC).

Alessie, R., and A. Kapteyn (1991), “Habit formation, interdependent preferences and demographic effects in the Almost Ideal Demand System”, Economic Journal 101:404–419.

Amaro de Matos, J.M.G., and J.F. Perez (1991), “Fluctuations in the Curie–Weiss version of the random field Ising model”, Journal of Statistical Physics 62(3/4):587–608.

Amemiya, T. (1985), Advanced Econometrics (Harvard University Press, Cambridge).

An, M., and N.M. Kiefer (1995), “Local externalities and societal adoption of technologies”, Journal of Evolutionary Economics 5:103–117.

Anderson, E. (1990), Streetwise (University of Chicago Press, Chicago).

Andreoni, J., and J.K. Scholz (1998), “An econometric analysis of charitable giving with interdependent preferences”, Economic Inquiry XXXVI:410–428.

Arthur, W.B. (1987), “Urban systems and historical path dependence”, in: R. Herman and J. Ausubel, eds., Urban Systems and Infrastructure (National Academy of Sciences/National Academy of Engineering, Washington, DC).

Arthur, W.B. (1989), “Increasing returns, competing technologies and lock-in by historical small events: the dynamics of allocation under increasing returns to scale”, Economic Journal 99:116–131.

Ash, R. (1972), Real Analysis and Probability (Academic Press, New York).

Averintsev, M. (1970), “On a method of describing discrete parameter fields”, Problems of Information Transmission 6:100–109.

Axtell, R., R. Axelrod, J. Epstein and M. Cohen (1996), “Aligning simulation models: a case study and results”, Computational and Mathematical Organization Theory 1(2):123–141.

Banerjee, A. (1992), “A simple model of herd behavior”, Quarterly Journal of Economics CVII:797–818.

Bauman, K., and L. Fisher (1986), “On the measurement of friend behavior in research on friend influence and selection: findings from longitudinal studies of adolescent smoking and drinking”, Journal of Youth and Adolescence 15:345–353.

Becker, G. (1973), “A theory of marriage: Part I” Journal of Political Economy 81:813–846.

Bell, A. (1995), “Dynamically interdependent preferences in a general equilibrium environment”, Mimeo (Department of Economics, Vanderbilt University).

Bénabou, R. (1993), “Workings of a city: location, education, and production”, Quarterly Journal of Economics CVIII:619–652.

Bénabou, R. (1996a), “Equity and efficiency in human capital investment: the local connection”, Review of Economic Studies 62:237–264.

Bénabou, R. (1996b), “Heterogeneity, stratification, and growth: macroeconomic effects of community structure”, American Economic Review 86:584–609.

Bernheim, D. (1994), “A theory of conformity”, Journal of Political Economy 5(102):841–877.

Bertrand, M., E. Luttmer and S. Mullainathan (1998), “Network effects and welfare cultures”, Working paper (Department of Economics, Harvard University).

Bikhchandani, S., D. Hirshleifer and I. Welch (1992), “A theory of fads, fashion, custom, and cultural exchange as information cascades”, Journal of Political Economy 100:992–1026.

Binder, M., and M.H. Pesaran (1998a), “Decision making in the presence of heterogeneous information and social interactions”, Working paper (Department of Economics, University of Maryland).

Binder, M., and M.H. Pesaran (1998b), “Life-cycle consumption under social interactions”, Working paper (Department of Economics, University of Maryland).

Blalock, H. (1984), “Contextual-effects models: theoretical and methodological issues”, Annual Review of Sociology 10:353–372.

Blau, D., and O. Duncan (1967), The American Occupational Structure (Wiley, New York).

Blume, L. (1993), “The statistical mechanics of strategic interaction”, Games and Economic Behavior 5:387–424.

Blume, L. (1995), “The statistical mechanics of best-response strategy revision”, Games and Economic Behavior 11:111–145.

Blume, L. (1997), “Population games”, in: W.B. Arthur, S. Durlauf and D. Lane, eds., The Economy as a Complex Evolving System II (Addison-Wesley, Redwood City).

Blume, L., and S. Durlauf (1998a), “The interactions-based approach to socioeconomic behavior”, Mimeo (Department of Economics, University of Wisconsin).

Blume, L., and S. Durlauf (1998b), “Equilibrium concepts for social interaction models”, Mimeo (Department of Economics, Cornell University).

Borjas, G. (1992), “Ethnic capital and intergenerational income mobility”, Quarterly Journal of Economics CVII:123–150.

Borjas, G. (1995), “Ethnicity, neighborhoods, and human-capital externalities”, American Economic Review 85:365–390.

Bretagnolle, J., and C. Huber-Carol (1988), “Effects of omitting covariates in Cox’s model for survival data”, Scandinavian Journal of Statistics 15:125–138.

Brewster, K. (1994a), “Race differences in sexual activity among adolescent women: the role of neighborhood characteristics”, American Sociological Review 59:408–424.

Brewster, K. (1994b), “Neighborhood context and the transition to sexual activity among black women”, Demography 31:603–614.

Brock, W. (1993), “Pathways to randomness in the economy: emergent nonlinearity and chaos in economics and finance”, Social Systems Research Institute Reprint no. 410 (Department of Economics, University of Wisconsin at Madison). Published in Estudios Economicos 8(1):3-55. Reprinted 1996, in: W. Dechert, ed., Chaos Theory in Economics: Methods, Models and Evidence (Edward Elgar, Cheltenham, UK).

Brock, W., and S. Durlauf (1995), “Discrete choice with social interactions”, Mimeo (Department of Economics, University of Wisconsin at Madison). Review of Economic Studies, forthcoming.

Brock, W., and S. Durlauf (1999), “A formal model of theory choice in science”, Economic Theory 14(1):113–130.

Brock, W., and C. Hommes (1998), “Rational routes to randomness”, Econometrica 65:1059–1096.

Brooks-Gunn, J., G. Duncan, P. Klebanov and N. Sealand (1993), “Do neighborhoods affect child and adolescent development?”, American Journal of Sociology 99:353–395.

Brown, B. (1990), “Peer groups and peer cultures”, in: S. Feldman and G. Elliott, eds., At the Threshold (Harvard University Press, Cambridge).

Brown, B., D. Clasen and S. Eicher (1986), “Perceptions of peer pressure, peer conformity dispositions, and self-reported behavior among adolescents”, Developmental Psychology 22:521–530.

Bryant, J. (1985), “A simple rational expectations Keynes-type model”, Quarterly Journal of Economics XCVIII:525–529.

Cameron, S., and J.J. Heckman (1998), “Life cycle schooling and dynamic selection bias: models and evidence for five cohorts of American males”, Journal of Political Economy 106:262–333.

Caplin, A., and J. Leahy (1994), “Business as usual, market crashes, and wisdom after the fact”, American Economic Review 84:548–565.

Case, A. (1992), “Neighborhood influence and technological change”, Regional Science and Urban Economics 22:491–508.

Case, A., and L. Katz (1991), “The company you keep: the effects of family and neighborhood on disadvantaged families”, Working Paper no. 3705 (National Bureau of Economic Research).

Casella, A., and J. Rauch (1997), “Anonymous market and group ties in international trade”, Working Paper no. 6186 (National Bureau of Economic Research).

Casella, A., and J. Rauch (1998), “Overcoming informational barriers to international resource allocation: prices and group ties”, Working Paper no. 6628 (National Bureau of Economic Research).

Chamberlain, G. (1984), “Panel data”, in: Z. Griliches and M. Intrilligator, eds., Handbook of Econometrics, Vol. 2 (North-Holland, Amsterdam).

Clark, A., and A. Oswald (1996), “Satisfaction and comparison income”, Journal of Public Economics 61:359–381.

Clark, A., and A. Oswald (1998), “Comparison-concave utility and following behaviour in social and economic settings”, Journal of Public Economics 70:133–155.

Cole, H., G. Mailath and A. Postlewaite (1992), “Social norms, savings behavior, and growth”, Journal of Political Economy 100:1092–1125.

Coleman, J. (1988), “Social capital in the creation of human capital”, American Journal of Sociology 94(supplement):S95-S120.

Coleman, J. (1990), Foundations of Social Theory (Harvard University Press, Cambridge).

Coleman, J., E. Campbell, J. Hobson, J. McPartland, A. Mood, F. Weinfeld and R. York (1966), Equality of Educational Opportunity (US Government Printing Office, Washington, DC).

Coleman, J., T. Hoffer and S. Kilgore (1982), High School Achievement: Public, Private and Catholic Schools Compared (Basic Books, New York).

Conley, T., and G. Topa (1999), “Socio-economic distance and spatial patterns in unemployment”, Mimeo (Department of Economics, New York University).

Conlisk, J. (1976), “Interactive Markov chains”, Journal of Mathematical Sociology 4:157–185.

Cooper, R., and A. John (1988), “Coordinating coordination failures in Keynesian models”, Quarterly Journal of Economics CIII:441–464.

Cooper, S. (1998), “A positive theory of income redistribution”, Journal of Economic Growth 3:171–195.

Corcoran, M., R. Gordon, D. Laren and G. Solon (1992), “The association between men’s economic status and their family and community origins”, Journal of Human Resources 27:575–601.

Crane, J. (1991a), “The epidemic theory of ghettos and neighborhood effects on dropping out and teenage childbearing”, American Journal of Sociology 96:1226–1259.

Crane, J. (1991b), “Effects of neighborhoods on dropping out of school and teenage childbearing”, in: C. Jencks and P. Peterson, eds., The Urban Underclass (Brookings Institution Press, Washington, DC).

Cross, P., and C.F. Manski (1999), “Regressions, long and short”, Mimeo (Department of Economics, Northwestern University).

Cutler, D., and E. Glaeser (1997), “Are ghettos good or bad?”, Quarterly Journal of Economics CXII:827–872.

Darrough, M., R. Pollak and T. Wales (1983), “Dynamic and stochastic structure: an analysis of three time series of household budget studies”, Review of Economics and Statistics 65:274–281.

Dasgupta, P., and P. David (1994), “Towards a new economics of science”, Research Policy 23:487–521.

David, P. (1985), “Clio and the economics of QWERTY”, American Economic Review 75:332–337.

David, P. (1998), “The collective cognitive performance of invisible colleges”, Working paper (Oxford University).

Duesenberry, J. (1949), Income, Saving, and the Theory of Consumer Behavior (Harvard University Press, Cambridge).

Duncan, G., and S. Raudenbusch (1998), “Neighborhoods and adolescent development: how can we assess the links?”, Working paper (Northwestern University).

Duneier, M., and H. Molotch (1999), “Talking city trouble: interactional vandalism, social inequality, and the “urban interaction problem”, Mimeo (University of Wisconsin). American Sociological Review, forthcoming.

Durlauf, S. (1993), “Nonergodic economic growth”, Review of Economic Studies 60:349–366.

Durlauf, S. (1996a), “A theory of persistent income inequality”, Journal of Economic Growth 1:75–93.

Durlauf, S. (1996b), “Neighborhood feedbacks, endogenous stratification, and income inequality”, in: W. Barnett, G. Gandolfo and C. Hillinger, eds., Dynamic Disequilibrium Modelling: Proceedings of the Ninth International Symposium on Economic Theory and Econometrics (Cambridge University Press, Cambridge).

Durlauf, S. (1997), "Statistical mechanics approaches to socioeconomic behavior", in: W.B. Arthur, S. Durlauf and D. Lane, eds., The Economy as a Complex Evolving System II (Addison-Wesley, Redwood City).

Durlauf, S., and D. Quah (1999), "The new empirics of economic growth", in: J. Taylor and M. Woodford, eds., Handbook of Macroeconomics (North-Holland, Amsterdam).

Elbers, C., and G. Ridder (1982), "True and spurious duration dependence: the identifiability of the proportional hazards model", Review of Economic Studies 49:403–410.

Ellis, R. (1985), Entropy, Large Deviations, and Statistical Mechanics (Springer, New York).

Emirbayer, M., and J. Goodwin (1997), "Network analysis, culture, and the problem of agency", American Journal of Sociology 99:1411–1454.

Epple, D., and T. Romer (1991), "Mobility and redistribution", Journal of Political Economy 99:828–858.

Epple, D., and H. Sieg (1999), "Estimating equilibrium models of local jurisdictions", Mimeo (Carnegie-Mellon University). Journal of Political Economy, forthcoming.

Epstein, J., and R. Axtell (1996), Growing Artificial Societies: Social Science from the Bottom Up (MIT Press, Cambridge).

Evans, W., and R. Schwab (1995), "Finishing high school and starting college: do Catholic schools make a difference?", Quarterly Journal of Economics CX:909–940.

Evans, W., W. Oates and R. Schwab (1992), "Measuring peer group effects: a study of teenage behavior", Journal of Political Economy 100:966–991.

Farrell, J., and G. Saloner (1985), "Standardization, compatibility and innovation", Rand Journal of Economics 16:70–83.

Fernandez, R., and R. Rogerson (1996), "Income distribution, communities, and the quality of public education", Quarterly Journal of Economics CXI:135–164.

Fingleton, B. (1999), "Economic geography with spatial econometrics: a 'third way' to analyze economic development and 'equilibrium' with application to the EU regions", Working Paper no. 99/21 (Department of Economics, European University Institute).

Föllmer, H. (1974), "Random economies with many interacting agents", Journal of Mathematical Economics 1:51–62.

Frank, R. (1985), Choosing the Right Pond (Oxford University Press, New York).

Freedman, D.A. (1991), "Statistical models and shoe leather", in: P. Marsden, ed., Sociological Methodology 1991 (Blackwell, Oxford).

Freedman, D.A., S. Klein, P. Sacks, J. Smythe and C. Everett (1991), "Ecological inference and voting rights (with discussion)", Evaluation Review 15:673–711.

Freedman, D.A., S. Klein, M. Ostland and M. Roberts (1998), "On 'solutions' to the ecological inference problem", Technical Report (Department of Statistics, UC Berkeley).

Glaeser, E., and J. Scheinkman (1998), "Measuring social interactions", Mimeo (Department of Economics, Harvard University).

Glaeser, E., B. Sacerdote and J. Scheinkman (1996), "Crime and social interactions", Quarterly Journal of Economics CXI:507–548.

Goering, J. (1996), Expanding Housing Choices for HUD-assisted Families: First Biennial Report on the Moving to Opportunity for Fair Housing Demonstration (Department of Housing and Urban Development, Office of Policy Development and Research, Washington, DC).

Goodman, L.A. (1953), "Ecological regression and the behavior of individuals", American Sociological Review 18:663–666.

Goolsbee, A., and P. Klenow (1998), "Evidence on network and learning externalities in the diffusion of home computers", Mimeo (Graduate School of Business, University of Chicago).

Gottfredson, M., and T. Hirschi (1990), A General Theory of Crime (Stanford University Press, Stanford).

Granovetter, M. (1995), Getting a Job, 2nd edition (University of Chicago Press, Chicago).

Granovetter, M., and R. Soong (1986), “Threshold models of interpersonal effects in consumer demand”, Journal of Economic Behavior and Organization 7:83–99.

Granovetter, M., and R. Soong (1988), “Threshold models of diversity: Chinese restaurants, residential segregation, and the spiral of silence”, in: C. Clogg, ed., Sociological Methodology 1988 (Blackwell, Oxford).

Hansen, L.P., and T. Sargent (1991), Rational Expectations Econometrics (Westview Press, Boulder).

Hauser, R. (1970), “Context and consex: a cautionary tale”, American Journal of Sociology 75:645–664.

Haveman, R., and B. Wolfe (1994), Succeeding Generations (Russell Sage Foundation, New York).

Heckman, J.J. (1979), “Sample selection bias as a specification error”, Econometrica 47:153–161.

Heckman, J.J. (1996), “Randomization as an instrumental variable”, Review of Economics and Statistics 73:336–340.

Heckman, J.J. (1997), “Instrumental variables”, Journal of Human Resources 32:441–462.

Heckman, J.J., and D. Neal (1996), “Coleman’s contributions to education: theory, research styles and empirical research”, in: J. Clark, ed., James S. Coleman (Falmer Press, Washington, DC).

Heckman, J.J., and R. Robb (1985), “Alternative methods for evaluating the impact of interventions”, in: J. Heckman and B. Singer, eds., Longitudinal Analysis of Labor Market Data (Cambridge University Press, New York).

Heckman, J.J., and B. Singer (1984a), “Econometric duration analysis”, Journal of Econometrics 24:63–132.

Heckman, J.J., and B. Singer (1984b), “A method for minimizing the impact of distributional assumptions in econometric models for duration data”, Econometrica 52:271–320.

Heckman, J.J., and B. Singer (1984c), “The identifiability of the proportional hazards model”, Review of Economic Studies 51:231–241.

Heckman, J.J., and B. Singer (1985), “Social science duration analysis”, in: J. Heckman and B. Singer, eds., Longitudinal Analysis of Labor Market Data (Cambridge University Press, New York).

Heckman, J.J., and J. Smith (1995), “Assessing the case for social experiments”, Journal of Economic Perspectives 9:83–110.

Heckman, J.J., H. Ichimura, J. Smith and P. Todd (1998a), “Characterizing selection bias using experimental data”, Econometrica 66:1017–1098.

Heckman, J.J., H. Ichimura and P. Todd (1998b), “Matching as an econometric evaluation estimator”, Review of Economic Studies 65:261–294.

Henderson, J.V., R. Mieszkowski and Y. Sauvageau (1978), “Peer group effects and educational production functions”, Journal of Public Economics, 10:97–106.

Hoffman, S., and R. Plotnick (1996), “The effect of neighborhood characteristics on young adult outcomes: alternative estimates”, Discussion Paper no. 1106-96 (Institute for Research on Poverty, University of Wisconsin).

Honoré, B.E., and E. Kyriazidou (1998), “Panel data discrete choice models with lagged dependent variables”, Mimeo (Department of Economics, Princeton University).

Ioannides, Y. (1990), “Trading uncertainty and market structure”, International Economic Review 31:619–638.

Ioannides, Y. (1997a), “Evolution of trading structures”, in: W.B. Arthur, S. Durlauf and D. Lane, eds., The Economy as a Complex Evolving System II (Addison-Wesley, Redwood City).

Ioannides, Y. (1997b), “Nonlinear neighborhood interactions and intergenerational human capital”, Working paper (Department of Economics, Tufts University).

Ioannides, Y. (1999), “Residential neighborhood effects”, Working paper (Department of Economics, Tufts University).

Jones, A. (1994), “Health, addiction, social interaction, and the decision to quit smoking”, Journal of Health Economics 13:93–110.

Jones, S. (1984), The Economics of Conformism (Basil Blackwell, Oxford).

Jovanovic, B. (1985), “Micro shocks and aggregate risk”, Quarterly Journal of Economics CII:395–409.

Jovanovic, B. (1989), "Observable implications of models with multiple equilibria", Econometrica 57:1431–1439.

Kac, M. (1968), "Mathematical Mechanisms of Phase Transitions", in: M. Chretien, E. Gross and S. Desar, eds., Statistical Physics: Phase Transitions and Superfluidity, Vol. 1 (Brandeis University Summer Institute in Theoretical Physics 1966).

Kapteyn, A., S. van de Geer, H. van de Stadt and T. Wansbeek (1997), "Interdependent preferences: An econometric analysis", Journal of Applied Econometrics 12:665–686.

Katz, L., J. Kling and J. Liebman (1997), "Moving to Opportunity in Boston: early impacts of a housing mobility program", Mimeo (Department of Economics, Harvard University).

Katz, M., and C. Shapiro (1986), "Technology adoption in the presence of network externalities", Journal of Political Economy 94:822–841.

Kelly, M. (1997), "The dynamics of Smithian growth", Quarterly Journal of Economics CXII:939–964.

King, G. (1997), A Solution to the Ecological Inference Problem (Princeton University Press, Princeton).

Kirman, A. (1983), "Communication in markets: a suggested approach", Economic Letters 12:1–5.

Kirman, A. (1997), "The economy as an interactive system", in: W.B. Arthur, S. Durlauf and D. Lane, eds., The Economy as a Complex Evolving System II (Addison-Wesley, Redwood City).

Kirman, A., C. Oddou and S. Weber (1986), "Stochastic communication and coalition formation", Econometrica 54:129–138.

Kitcher, P. (1993), The Advancement of Science (Oxford University Press, Oxford).

Kollman, K., J. Miller and S. Page (1992), "Adaptive parties in spatial elections", American Political Science Review 86:929–937.

Kollman, K., J. Miller and S. Page (1997a), "Political parties and electoral Landscapes", British Journal of Political Science, forthcoming.

Kollman, K., J. Miller and S. Page (1997b), "Computational political economy", in: W.B. Arthur, S. Durlauf and D. Lane, eds., The Economy as a Complex Evolving System II (Addison-Wesley, Redwood City).

Kremer, M. (1997), "How much does sorting increase inequality?", Quarterly Journal of Economics CXII:115–140.

Krosnick, J., and C. Judd (1982), "Transitions in social influence in adolescence: who induces cigarette smoking", Developmental Psychology 81:359–368.

Krugman, P. (1996), The Self-Organizing Economy (Blackwell, Oxford).

Kuhn, T. (1970), The Structure of Scientific Revolutions, 2nd edition (University of Chicago Press, Chicago).

Kyriazidou, E. (1997a), "Estimation of a panel data sample selection model", Econometrica 65: 1335–1364.

Kyriazidou, E. (1997b), "Estimation of dynamic panel data sample selection model", Mimeo (University of Chicago).

Labov, W. (1972a), Sociolinguistic Patterns (University of Pennsylvania Press, Philadelphia).

Labov, W. (1972b), Language in the Inner City (University of Pennsylvania Press, Philadelphia).

Ladd, H., and J. Ludwig (1998), "The effects of MTO on educational opportunities in Baltimore: early evidence", Working paper (Joint Center for Poverty Research, Northwestern University).

Lancaster, T. (1990), The Analysis of Transition Data (Cambridge University Press, New York).

Leamer, E.E. (1983), "Let's take the con out of econometrics", American Economic Review 73:31–43.

Legros, P., and A. Newman (1997), "Matching in perfect and imperfect worlds", Mimeo (Department of Economics, Columbia University).

Lindbeck, A., S. Nyberg and J. Weibull (1999), "Social norms and economic incentives in the welfare state", Quarterly Journal of Economics CXIV:1–36.

Loury, G. (1977), "A dynamic theory of racial income differences", in: P. Wallace and A. LaMond, eds., Women, Minorities, and Employment Discrimination (Lexington Books. Lexington).

Lukacs, E. (1975), Stochastic Convergence (Academic Press, New York).

Manski, C.F. (1975), "The maximum score estimation of the stochastic utility model of choice", Journal of Econometrics 3:205–228.

Manski, C.F. (1985), "Semiparametric estimation of discrete response: asymptotic properties of the maximum score estimator", Journal of Econometrics 27:313–333.

Manski, C.F. (1988), "Identification of binary response models", Journal of the American Statistical Association 83:729–738.

Manski, C.F. (1993a), "Identification problems in the social sciences", in: P. Marsden ed., Sociological Methodology 1993 (Blackwell, Oxford).

Manski, C.F. (1993b), "Identification of endogenous social effects: the reflection problem", Review of Economic Studies 60:531–542.

Manski, C.F. (1995), Identification Problems in the Social Sciences (Harvard University Press, Cambridge).

Manski, C.F. (1997), "Identification of anonymous endogenous interactions", in: W.B. Arthur, S. Durlauf and D. Lane, eds., The Economy as a Complex Evolving System II (Addison-Wesley, Redwood City).

Manski, C.F., and J. Pepper (1998), "Monotone instrumental variables: with an application to the returns on schooling, Mimeo (Northwestern University). Econometrica, forthcoming.

McFadden, D. (1974), "Conditional logit analysis of qualitative choice behavior", in: P. Zarembka, ed., Frontiers in Econometrics (Academic Press, New York).

McFadden, D. (1981), "Econometrics models of probabilistic choice", in: C. Manski and D. McFadden, eds., Structural Analysis of Discrete Data with Econometric Applications (MIT Press, Cambridge, MA).

McFadden, D. (1984), "Econometric analysis of qualitative response models", in: Z. Griliches and M. Intrilligator, eds., Handbook of Econometrics, Vol. 2 (North-Holland, Amsterdam).

McManus, D. (1992), "How common is identification in parametric models?", Journal of Econometrics 53:5–23.

Milgrom, P., and D.J. Roberts (1990), "Rationalizability, learning, and equilibrium in games with strategic complementarities", Econometrica 58:1255–1277.

Miyao, T. (1978), "A probabilistic model of location choice with neighborhood effects", Journal of Economic Theory 19:34–58.

Moffitt, R. (1998), "Policy interventions, low-level equilibria, and social interactions", Mimeo (Department of Economics, Johns Hopkins University).

Montgomery, J. (1991), "Social networks and labor-market analysis: towards an economic analysis", American Economic Review 81:1408–1418.

Montgomery, J. (1992), "Social networks and persistent inequality in labor markets", Mimeo (Department of Economics, Northwestern University).

Morris, S. (1998), "Contagion", Mimeo (Department of Economics, Yale University). Review of Economic Studies, forthcoming.

Munshi, K., and J. Myaux (1998), "Social effects in the demographic transition: evidence from Matlab, Bangladesh", Mimeo (Department of Economics, University of Pennsylvania).

Murray, J. (1984), Asymptotic Analysis (Springer, New York).

Neal, D. (1997), "The effects of Catholic secondary schooling on educational achievement", Journal of Labor Economics 15:98–123.

Nechyba, T. (1996), "Social approval, values and AFDC: a re-examination of the illegitimacy debate", Mimeo (Department of Economics, Stanford University).

Newey, W.K., and D. McFadden (1994), "Large sample estimators and hypothesis testing", in: R. Engle and D. McFadden, eds., Handbook of Econometrics, Vol. 4 (North-Holland, Amsterdam) pp. 2113–2245.

Oomes, N. (1998), "Market failures in the economics of science: barriers to entry, increasing returns, and lock-in by historical events", Mimeo (Department of Economics, University of Wisconsin at Madison).

Page, S. (1997), "Network structure matters", Mimeo (Department of Economics, University of Iowa).

Pakes, A. (1999), "A framework for applied dynamic analysis in I.O.", Mimeo (Department of Economics, Harvard University).

Plotnick, R., and S. Hoffman (1996), "The effect of neighborhood characteristics on young adult outcomes: alternative estimates." Research discussion paper (Institute of Poverty, Madison, WI) pp. 1106–1196.

Popkin, S., J. Rosenbaum and P. Meaden (1993), "Labor market experiences of low-income black women in middle-class suburbs: evidence from a survey of Gautreaux program participants", Journal of Policy Analysis and Management 12:556–574.

Postlewaite, A. (1997), "The social basis of interdependent preferences", Mimeo (University of Pennsylvania).

Raftery, A.E. (1995), "Bayesian model selection in social science research", in: P. Marsden, ed., Sociological Methodology 1995 (Blackwell, Oxford).

Raftery, A.E., A.D. Madigan and J.A. Hoeting (1997), "Bayesian model averaging for linear regression models", Journal of the American Statistical Association 92:179–191.

Rivkin, S. (1997), "The estimation of peer group effects", Mimeo (Department of Economics, Amherst College).

Romer, D. (1993), "Rational asset-price movements without news", American Economic Review 83: 1112–1130.

Rosenbaum, J. (1995), "Changing the geography of opportunity by expanding residential choice: lessons from the Gautreaux program", Housing Policy Debate 6:231–269.

Rosenbaum, J., and S. Popkin (1991), "Employment and earnings of low-income blacks who move to middle class suburbs", in: C. Jencks and P. Peterson, eds., The Urban Underclass (Brookings Institution Press, Washington, DC).

Rosenbaum, J., S. DeLuca and S. Miller (1999), "The long-term effects of residential mobility on AFDC receipt: studying the Gautreaux program with administrative data", Mimeo (Northwestern University).

Rosser, J.B. (1999), "On the complexities of complex economic dynamics", Journal of Economic Perspectives 13(4):169–192.

Rothenberg, T. (1971), "Identification in parametric models", Econometrica 39:577–591.

Rouse, C.E. (1998a), "Private school vouchers and student achievement: an evaluation of the Milwaukee Parental Choice Program", Quarterly Journal of Economics CXIII:553–602.

Rouse, C.E. (1998b), "Schools and student achievement: more evidence from the Milwaukee Parental Choice Program", Working paper no. 396 (Industrial Relations Section, Princeton University).

Sah, R. (1991), "Social osmosis and crime", Journal of Political Economy 99:1272–1295.

Sampson, R., and J. Laub (1995), Crime in the Making (Harvard University Press, Cambridge).

Sampson, R., S. Raudenbusch and F. Earls (1997), "Neighborhoods and violent crime: A study of collective efficacy", Science 277:918–924.

Sattinger, M. (1975), "Comparative advantage and the distribution of earnings", Econometrica 43: 455–468.

Schelling, T. (1971), "Dynamic models of segregation", Journal of Mathematical Sociology 1:143–186.

Sims, C.A. (1980), "Macroeconomics and reality", Econometrica 48:1–48.

Slavin, R. (1990), "Achievement effects of ability grouping in secondary schools: a best-evidence synthesis", Review of Educational Research 60:471–499.

Solon, G., M. Page and G. Duncan (1999), "Correlations between neighboring children in their subsequent educational attainment", Mimeo (University of Michigan).

Spitzer, F. (1971), "Markov random fields and Gibbs ensembles", American Mathematical Monthly 78:142–154.

Steinberg, L., B. Brown and S. Dornbusch (1996), Beyond the Classroom (Simon and Schuster, New York).

Streufert, P. (1991), "The effect of underclass isolation on schooling choice", Mimeo (University of Wisconsin at Madison).

Sucoff, C., and D. Upchurch (1998), “Neighborhood context and the risk of childbearing among metropolitan-area black adolescents”, American Sociological Review 63:571–585.

Topa, G. (1997), “Social interactions, local spillovers and unemployment”, Mimeo (New York University). Review of Economic Studies, forthcoming.

Verbrugge, R. (1999), “A framework for studying economic interactions (with applications to corruption and business cycles)”, Mimeo (Department of Economics, Virginia Polytechnic Institute).

Wallis, K. (1980), “Econometric implications of the rational expectations hypothesis”, Econometrica 48:49–73.

Weinberg, B., P. Reagan and J. Yankow (1999), “Do neighborhoods affect work behavior? Evidence from the NLSY79”, Mimeo (Department of Economics, Ohio State University).

Witte, J. (1997), “Achievement effects of the Milwaukee voucher program”, Mimeo (Department of Political Science, University of Wisconsin).

Yinger, J. (1995), Closed Doors, Opportunities Lost (Russell Sage Foundation, New York).

Young, H.P. (1998), Individual Strategy and Social Structure (Princeton University Press, Princeton).

Zax, J., and D. Rees (1998), “Environment, ability, effort, and earnings”, Working paper no. 9801 (Center for Research on Economic and Social Policy, University of Colorado at Denver).

Chapter 55

DURATION MODELS: SPECIFICATION, IDENTIFICATION AND MULTIPLE DURATIONS

GERARD J. VAN DEN BERG*

*Department of Economics, Free University Amsterdam**, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands; CEPR; and Tinbergen Institute*

Contents

Abstract 3383
Keywords 3383
1. Introduction 3384
2. Basic concepts and notation 3387
3. Some structural models of durations 3389
3.1. Standard search model 3389
3.1.1. Stationarity 3389
3.1.2. Nonstationarity without anticipation 3391
3.1.3. Nonstationarity with anticipation 3391
3.2. Repeated-search model 3393
4. The Mixed Proportional Hazard model 3394
4.1. Definition 3394
4.2. Time-varying explanatory variables 3397
4.3. Theoretical justification 3400
5. Identification of the MPH model with single-spell data 3405
5.1. Some implications of the MPH model specification 3405
5.2. Identification results 3408
5.3. Interaction between duration and explanatory variables in the observed hazard 3412
5.4. The sign of the interaction 3414
5.5. Specification of the unobserved heterogeneity distribution 3418
5.5.1. Discrete distributions 3418
5.5.2. Gamma distributions 3419
5.5.3. Suggestions from economic theory 3420
5.5.3.1. Suggestions from equilibrium search models 3420

* Thanks to Tony Lancaster, Geert Ridder, Jean-Marc Robin, Jim Heckman, Zvi Eckstein and Jaap Abbring for their very useful suggestions. Also thanks to colleagues at the Free University Amsterdam for stimulating discussions.

** e-mail: gberg@econ.vu.nl.

Handbook of Econometrics, Volume 5, Edited by J.J. Heckman and E. Leamer

5.5.3.2. Suggestions from on-the-job search models 3421
5.6. Effects of misspecification of functional forms 3422
6. The MPH model with multi-spell data 3426
6.1. Multi-spell data 3426
6.2. Identification results 3427
7. An informal classification of reduced-form multiple-duration models 3431
8. The Multivariate Mixed Proportional Hazard model 3437
8.1. Definition 3437
8.2. Identification results 3438
8.2.1. Competing risks 3439
8.2.2. Successive durations 3441
8.3. Specification of the bivariate unobserved heterogeneity distribution 3442
8.3.1. Dimensionality 3442
8.3.2. The dependence between the durations 3444
9. Causal duration effects and selectivity 3447
9.1. Lagged endogenous durations 3447
9.2. Endogenous shocks 3448
10. Conclusions and recommendations 3449
References 3453

Abstract

Since the early 1980s, the econometric analysis of duration variables has become widespread. This chapter provides an overview of duration analysis, with an emphasis on the specification and identification of duration models, and with special attention to models for multiple durations. Most of the chapter deals with so-called reduced-form duration models, notably the popular Mixed Proportional Hazard (MPH) model and its multivariate extensions. The MPH model is often used to describe the relation between the empirical exit rate and "background variables" in a concise way. However, since the applications usually interpret the results in terms of some economic-theoretical model, we examine to what extent the deep structural parameters of some important theoretical models can be related to reduced-form parameters. We subsequently examine the specification and identification of the MPH model in great detail, we provide intuition on what drives identification, and we infer to what extent biases may occur because of misspecifications. This examination is carried out separately for the case of single-spell data and the case of multi-spell data. We also compare different functional forms for the unobserved heterogeneity distribution.

Next, we examine models for multiple durations. In the applied econometric literature on the estimation of multiple-duration models, the range of different models is actually not very large. Typically, the models allow for dependence between the duration variables by way of their unobserved determinants, with each single duration following its own MPH model. In addition to this, the model may allow for an interesting "causal" effect of one duration on the other, as motivated by an underlying economic theory. For all these models we examine the conditions for identification. Some of these are intimately linked to particular estimation strategies. The multiple-duration model where the marginal duration distributions each satisfy an MPH specification, and the durations can only be dependent by way of their unobserved determinants, is called the Multivariate Mixed Proportional Hazard (MMPH) model. For this model, we address the issue of the dimensionality of the heterogeneity distribution and we compare the flexibility of different parametric heterogeneity distributions.

On a number of occasions, we incorporate recent insights from the biostatistical literature on duration analysis, and we contrast points of view in this literature to those in the econometric literature. Finally, throughout the chapter, we discuss the importance of the possible collection of additional data.

Keywords

duration analysis, duration model, hazard rate, hazard function, transition rate, unemployment duration, identification, proportional hazard, failure time, failure rate, multiple spells, duration dependence, unobserved heterogeneity, treatment effect, shocks

JEL classification: C41, C51, J64, C24, C20, C23, C34, J63, C33

1. Introduction

Duration analysis is a core subject of econometrics. Since the early 1980s, the empirical analysis of duration variables has become widespread. There are a number of distinct reasons for this development. First of all, many types of behavior over time tend increasingly to be regarded as movements at random intervals from one state to another. Examples include movements by individuals between the labor market states of employment, unemployment and nonparticipation, and movements between different types of marital status. This development reflects the fact that dynamic aspects of economic behavior have become more important in economic theories, and that in these theories the arrival of new information (and thus the change in behavior in response to this) occurs at random intervals. Secondly, longitudinal data covering more than just one spell per respondent are widely available in labor economics, as well as in demography and medical science. Applications of duration analysis include, in labor economics, the duration of unemployment and the duration of jobs [see e.g., the survey by Devine and Kiefer (1991)], strike durations [e.g., Kennan (1985)], and the duration of training programs [Bonnal, Fougère and Sérandon (1997)]. In business economics, duration models have been used to study the duration until a major investment [e.g., Anti Nilsen and Schiantarelli (1998)]. In population economics, duration analysis has been applied to study marriage durations [Lillard (1993)], the duration until the birth of a child [Heckman and Walker (1990)], and the duration until death. In econometric analyses dealing with selective observation, duration models have been used to study the duration of panel survey participation [e.g., Van den Berg and Lindeboom (1998)]. In marketing, duration models have been used to study household purchase timing [e.g., Vilcassim and Jain (1991)], in consumer economics to study the duration until purchase of a durable or storable product [Antonides (1988), Boizot, Robin and Visser (1997)], and in migration economics to study the duration until return migration [e.g., Lindstrom (1996)]. Recently, duration models have been applied in areas in economics where the unit under consideration is not an individual or firm. For example duration models have been used in macro economics to study the duration of business cycles [e.g., Diebold and Rudebusch (1990)], in finance to study the duration between stock-market share transactions [Engle and Russell (1998)], in political economics to study the duration of wars [see Horvath (1968)], and in industrial organization to study the duration of a patent [Pakes and Schankerman (1984)].

This chapter presents an overview of duration analysis. A substantial part of the chapter deals with so-called reduced-form duration models, notably the famous Mixed Proportional Hazard (MPH) model. This model expresses the exit rate to a destination state as a rather simple function of observed and unobserved explanatory variables and the elapsed duration in the current state. This model and its special cases, most notably the Proportional Hazard (PH) model, have been used in hundreds of empirical studies [see e.g., Devine and Kiefer (1991) for references in micro labor economics]. Parametric versions of the model are included in statistical packages like STATA, SAS, S-PLUS and SPSS [see Pelz and Klein (1996) for a comparison of some packages]. We examine the specification and identification of the MPH model in detail, and we infer to what extent biases may occur because of misspecifications.

The MPH model is often used to describe the relation between the empirical exit rate and "background variables" in a concise way, and to provide estimates of the effect of an explanatory variable on the duration variable. However, since the applications usually interpret the results in terms of some economic-theoretical model, it is important to examine to what extent the deep structural parameters of this theoretical model can be related to the reduced-form parameters. As we shall see, economic theory in general does not lead to a "proportional" specification as in the MPH duration model, and this complicates the interpretation of the reduced-form estimates.

Recently, the empirical analysis of multiple durations has become widespread. In many cases it is simply a necessity to address the issue of whether different durations (given the observed explanatory variables) are not independently distributed. For example, if the duration data are censored then it matters for empirical inference how the time until censoring is related to the duration of interest. More generally, if a spell under observation can terminate in a number of different ways ("competing risks") then it matters whether the latent durations to the different destinations are related. As we shall see, economic theory often predicts that such durations are related. In fact, the issue of whether different durations are related is often an important question in its own right. Because of this, current econometric research often involves the simultaneous analysis of multiple observed spells of the same type of duration for a given individual, or multiple observed spells of different types of durations for a given individual. For example, it may involve simultaneous and consecutive durations in labor market states and marital states. It may also involve the analysis of treatment effects on a duration variable, if the duration until treatment (or the duration of the treatment) is stochastic. In this chapter we therefore pay special attention to the analysis of multiple durations. We examine different types of relations between duration variables, as motivated by economic theory. We then examine the way in which they can be incorporated in multivariate extensions[1] of the MPH model, and we discuss identification of the determinants of these multivariate models as well as identification of deep structural parameters. For the case where the dependence runs by way of related unobserved explanatory variables (in which case we call the model a multivariate MPH (MMPH) model), we compare different parametric heterogeneity distributions. One of the main conclusions of the sections on multiple-duration models is that, in microeconometric research involving self-selection, duration data are much more informative than binary data. This is important because economic theory generally predicts the absence of exclusion restrictions based

[1] In this chapter, "multivariate" refers to multiple durations and not to multiple explanatory variables.

on characteristics of the individual under consideration, so that these can not be used for identification.

So far, we have been vague on the meaning of notions like “state”, “duration”, “exit rate”, and “explanatory variable”. In Section 2 we provide some formal definitions. We stress that the economic meaning of these notions is entirely context-dependent: what distinguishes states or transitions in one study may not be relevant in another study. Throughout the chapter we will be concerned with the economic insights that can be obtained from duration analysis. For that reason we outline in Section 3 some motivating underlying economic models for durations. In particular, we examine search models of individual labor market behavior. After these preparatory sections we examine the MPH model in Sections 4 and 5. Section 6 deals with the identification of the MPH model in case the data provide durations of multiple spells in a given state for a given individual. Such data are called multi-spell data. Again, the meaning of these notions is rather vague at this stage. Basically, the idea is that the data provide multiple independent drawings from the individual-specific duration distribution. Sections 7–9 deal with multiple-duration models in general. These constitute a very broad class of models, and they include, as a special case, the model of Section 6 with durations of multiple spells in a given state for a given individual. Section 10 concludes and provides recommendations on empirical approaches.

Throughout the chapter, time is taken to be continuous[2]. When specifying a duration distribution, the point of departure will invariably be the exit rate or hazard rate (this is motivated in Section 2). This implies that we do not focus on so-called Accelerated Failure Time models [see e.g., Kalbfleisch and Prentice (1980)], which enjoy some popularity outside economics. At times, though, we compare the latter models to models that are based on a specification of the hazard rate.

In this chapter we do not focus on estimation methods or specification tests. Applied studies generally use well-established estimation methods like Maximum Likelihood, Cox Partial Likelihood, Conditional Likelihood, or nonparametric methods. The book by Lancaster (1990), which is the most comprehensive volume on econometric duration analysis so far, provides an excellent survey on estimation methods and specification tests for MPH models in econometrics. Andersen et al. (1993) survey the literature on the modern statistical foundations. Kiefer (1988) and Yamaguchi (1991) lucidly explain the basics of the empirical analysis of duration models. Finally, the survey by Neumann (1997) discusses specification tests as well, and also pays attention to the estimation of structural (search) models.

[2] See Meyer (1995) for a survey of discrete-time reduced-form duration models. These models include continuous-time models where time is aggregated into intervals of unit length, as well as models where time is genuinely discrete.

2. Basic concepts and notation

Consider the spells experienced by certain subjects in a certain state. The duration of the spell is stochastic and is denoted by T, and realizations of T are denoted by t[3]. The cumulative distribution function of T is denoted by F, so $F(t) = \Pr(T \leqslant t)$, with $F(0) = 0$. The *survivor function* of T is defined as one minus the distribution function and is denoted by $\overline{F}$, so

$$\overline{F}(t) = 1 - F(t).$$

As noted in the introduction, we restrict attention to continuous random variables T, and we denote a probability density function of T by f. In fact, F, $\overline{F}$, and f will be used as generic symbols for cumulative distribution functions, survivor functions, and probability density functions, respectively, and their arguments make clear which random variable is considered.

In a discrete-time setting, the *hazard function* of T at t is defined as the probability that the spell is completed at t given that it has not been completed before t, as a function of t. With T continuous, we define the hazard function as

$$\theta(t) = \lim_{\mathrm{d}t \downarrow 0} \frac{\Pr(T \in [t, t + \mathrm{d}t) | T \geqslant t)}{\mathrm{d}t}.$$

So, somewhat loosely, the hazard function is the rate at which the spell is completed at t given that it has not been completed before, as a function of t. The value of the hazard function (for a particular t, or for arbitrary t) is called the “hazard rate” or simply “the hazard”. It is also called the “exit rate” to stress the fact that completion of the spell is equivalent to exit out of the state of interest. Again, we use θ as a generic symbol for a hazard, and its argument makes clear which random variable is considered. The hazard function $\theta(t)$ is said to be duration dependent if its value changes over t. Positive (negative) duration dependence means that $\theta(t)$ increases (decreases).

The hazard function provides a full characterization of the distribution of T, just like the distribution function, the survivor function, and the density function. All of these can be expressed in terms of one another. For F, $\overline{F}$, and f this is well known. Concerning θ, the following relations (which are easy to derive) express θ in terms of the other functions, and vice versa,

$$\theta(t) = \frac{f(t)}{1 - F(t)},$$

$$\overline{F}(t) = \exp\left(-\int_0^t \theta(u)\,\mathrm{d}u\right) \qquad t \geqslant 0. \tag{1}$$

The hazard function is the focal point of econometric duration models. That is, properties of the distribution of T are generally discussed in terms of properties of θ.

[3] Throughout most of the chapter, we use t to denote the random variable as well as its realization. This abusive notation has become common in duration analysis because it allows for concise formulations that are generally unambiguous.

There are two major reasons for this. First, and most importantly, this approach is dictated by economic theory. In general, theories that aim at explaining durations focus on the rate at which the subject leaves the state at duration t given that he has not done so yet. In particular, they explain the hazard at t in terms of external conditions at t as well as the underlying economic behavior of the subjects that are still in the state at t. Theoretical predictions about a duration distribution thus run by way of the hazard of that distribution. It is obvious that if the completion of a spell is at least partly affected by external conditions that change over time (e.g., due to external shocks), and if one attempts to describe behavior of the subject over time in a changing environment, then it is easier to think about the rate of leaving at t given that one has not done so than to focus on the unconditional rate of leaving at t. In the next section we provide some examples of such theories.

It is often stated that a major advantage of using the hazard function as a basic building block of the model is that it facilitates the inclusion of time-varying covariates. This is, of course, part of the argument of the previous paragraph; it reformulates the issue from the point of view of a builder of reduced-form models.

The second major advantage of using the hazard function as the basic building block of the model is entirely practical. Real-life duration data are often subject to censoring of high durations. In that case it does not make sense to model the duration distribution for those high durations.

Whereas the hazard function is the focal point of model building in duration analysis, the mean of the endogenous variable is the focal point in regression analysis. On some occasions in the chapter we compare duration models to regression models. For future reference it is useful to present the equation below. This equation follows directly from the fundamental result that the integrated hazard function $\int_0^t \theta(u)\,\mathrm{d}u$ has an exponential distribution[4] with parameter 1.

$$\log \int_0^t \theta(u)\,\mathrm{d}u = \varepsilon. \tag{3}$$

Here, ε has an Extreme Value – Type I (EV1) distribution. This distribution does not have any unknown parameters; its density equals

Extreme Value – Type 1 distribution:

$$f(\varepsilon) = \mathrm{e}^{\varepsilon} \cdot \mathrm{e}^{-\exp(\varepsilon)}, \quad \text{for all } -\infty < \varepsilon < \infty.$$

Equation (3) therefore again shows that once the hazard function is completely specified, then so is the duration distribution. Note that the transformation of t on

[4] Family of exponential distributions:

$$f(t) = \vartheta \mathrm{e}^{-\vartheta t} \quad \text{for all } t \geq 0, \text{ with } \vartheta > 0. \tag{2}$$

the left-hand side of Equation (3) can be interpreted as a particular change in the time measurement scale. The equation states that after this transformation, the only variation left in the duration concerns the purely random variation that is unrelated to the determinants of $\theta(t)$. Note that if one specifies a model for $\theta(t)$ then a natural model specification test follows from a comparison of the empirical distribution of the estimated left-hand side of Equation (3) to the distribution of ε [see Lancaster (1990)].

3. Some structural models of durations

In this section we briefly discuss some economic-theoretical models that predict distributions of duration variables. These theoretical models have been structurally estimated using data on such duration variables, and they have been used to interpret estimates of reduced-form duration models. The common feature of the models is that they are search models, which describe the duration until an event as the outcome of a decision on the optimal moment of stopping the search for something desirable[5]. For expositional reasons we phrase the models in terms of search for jobs by individual agents on the labor market (although they are applicable to many other types of search). Job search models have been very popular as explanatory theoretical frameworks for reduced-form econometric duration analyses [see Devine and Kiefer (1991)].

3.1. Standard search model

3.1.1. Stationarity

In this subsection we consider the prototype job search model for the behavior of unemployed workers. Here, the duration variable of interest is the unemployment duration. Since this model has been discussed extensively many times [e.g., Mortensen (1986)], the present exposition is brief.

The model aims to describe the behavior of unemployed individuals in a dynamic and uncertain environment. Job offers arrive at random intervals following a Poisson process with arrival rate λ. A job offer is a random drawing (without recall) from a wage offer distribution with distribution function $F(w)$[6]. It is assumed that all jobs are full-time jobs. Every time an offer arrives, the decision has to be made whether to accept the offer or reject it and search further. Once a job is accepted it will be

[5] There are many other theoretical models that give rise to duration distributions. Examples are learning models [see e.g., Jovanovic (1984)] and dynamic discrete choice models [see e.g., Rust (1994) for a survey]. The latter can be considered as generalizations of basic search models although they are necessarily in discrete time; as such they give rise to discrete duration distributions. These models may also be used to explain multiple durations for a given subject [see e.g., Van der Klaauw (1996)].

[6] Note that F here denotes a distribution of wage offers rather than a duration distribution.

held forever at the same wage, so job-to-job transitions are excluded. It is assumed that individuals know λ and F but that they do not know in advance when job offers arrive and what wages are associated with them. During the spell of unemployment a benefit b is received. Unemployed individuals aim at maximization of their own expected present value of income over an infinite horizon. The subjective rate of discount is denoted by ρ.

The variables λ, w, b and ρ are measured per unit time period. It is assumed that the model is stationary. This means that λ, F, b and ρ are assumed to be constant, and, in particular, independent of unemployment duration and calendar time and independent of all events during unemployment. To ensure that attention is restricted to economically meaningful cases, and to guarantee the existence of the optimal strategy, we assume that $0 < \lambda, \mathrm{E}_F(w), b, \rho < \infty$. For ease of exposition we take F to be continuous.

Let R denote the expected present value of search when following the optimal strategy. Because of the stationarity assumption and the infinite-horizon assumption, the unemployed individual's perception of the future is independent of time or unemployment duration, so the optimal strategy is constant during the spell of unemployment and R does not depend on the elapsed unemployment duration t. It is well known [see e.g., Mortensen (1986)] that there is a unique solution to the Bellman equation for R, satisfying

$$\rho R = b + \lambda \mathrm{E}_w \max\{0, \frac{w}{\rho} - R\}. \tag{4}$$

In this equation, the expectation is taken over the wage offer distribution F. Equation (4) has a familiar structure [see e.g., Pissarides (1990)]. The return of the asset R in a small interval around t equals the sum of the instantaneous utility flow in this interval, and the expected excess value of finding a job in this interval. When an offer of w arrives at t then there are two options: (i) to reject it (excess value zero), and (ii) to accept it (excess value $w/\rho - R$). It is clear that the optimal policy is to choose option (ii) iff $w > \rho R$. Therefore, the optimal strategy of the worker can be characterized by a reservation wage ϕ: a job offer is acceptable iff its wage exceeds ϕ, with $\phi = \rho R$. Using Equation (4), ϕ can be expressed in terms of the model determinants,

$$\phi = b + \frac{\lambda}{\rho} \int_{\phi}^{\infty} \overline{F}(w)\, \mathrm{d}w.$$

Note that this equation has a unique solution for ϕ.

The hazard (or exit rate out of unemployment, or transition rate from unemployment into employment) θ equals the product of the job offer arrival rate and the conditional probability of accepting a job offer,

$$\theta = \lambda \overline{F}(\phi).$$

As a result of the stationarity assumption, θ does not depend on the elapsed duration of unemployment. Consequently, the duration of unemployment t has an exponential distribution (see Equation 2) with parameter θ.

Versions of this model have been structurally estimated with individual data on unemployment durations and wages. “Structural” here means that the theoretical framework is assumed to describe the empirical distribution of durations and wages. This enables estimation of the determinants λ, F, ... of individual behavior. See Yoon (1981), Flinn and Heckman (1982a), Narendranathan and Nickell (1985) and Van den Berg (1990b) for examples of this, and Wolpin (1995) for a survey.

3.1.2. Nonstationarity without anticipation

The stationarity assumption made in the previous subsection is often unrealistic. The values of the structural determinants may change because of duration dependence of the amount of unemployment benefits, a stigma effect of being long-term unemployed, policy changes, or business cycle effects. Sooner or later these features of the labor market and personal characteristics of job searchers are recognized and used in determining the optimal strategy. So, generally, the optimal strategy is not constant in case of nonstationarity.

To proceed, assume that the individual’s search environment is subject to unanticipated changes in the values of the structural determinants. Thus, the values of these determinants may change over the duration, but the individual always thinks that they will remain constant at their current values. This might be a reasonable assumption in case of a change in λ that is due to a random macroeconomic shock, or in case of a change in b that is due to a sudden change in the benefits system.

By exploiting the analogy to the stationary model, we obtain the following equations for the reservation wage function $\phi(t)$, giving the reservation wage at time t, and the hazard function $\theta(t)$,

$$\phi(t) = b(t) + \frac{\lambda(t)}{\rho(t)} \int_{\phi(t)}^{\infty} \overline{F}(w|t)\,\mathrm{d}w, \qquad \theta(t) = \lambda(t)\overline{F}(\phi(t)|t),$$

where $F(w|t)$ denotes the wage offer distribution at time t (so it should not be interpreted as a distribution conditional on the realization of a random duration variable). In general, $\theta(t)$ varies with t. The distribution function for the duration of unemployment subsequently follows from Equation (1). See Narendranathan (1993) for a structural empirical analysis of a nonstationary model without anticipation.

3.1.3. Nonstationarity with anticipation

In many cases it is not realistic to assume that individuals do not anticipate changes in the values of λ, F and b. In this subsection we consider nonstationarity with anticipation, along the lines of Van den Berg (1990a)[7]. The structural determinants λ, F and b are allowed to vary over the duration t in a deterministic way (so dependence on past offer arrival times or wage levels associated with rejected offers is ruled out). This entails that the process with which job offers arrive is a non-homogeneous Poisson process. We assume that job searchers have perfect foresight in the sense that they correctly anticipate changes in the values of λ, F and b. In other words, we expect people to know how these are related to t. As usual, individuals do not know in advance when job offers arrive, or which w are associated with them. Finally, we assume that λ, F and b are constant for all sufficiently high t. The latter implies that the optimal strategy is also constant for sufficiently high t.

Let $R(t)$ denote the expected present value of search if unemployment duration equals t, when following the optimal strategy. Under regularity conditions, there is a unique continuous solution to the Bellman equation for $R(t)$, satisfying

$$\rho R(t) = \frac{\mathrm{d}R(t)}{\mathrm{d}t} + b(t) + \lambda(t) \cdot \mathrm{E}_{w|t} \max\{0, \frac{w}{\rho} - R(t)\},$$

at points at which $R(t)$ is differentiable in t, where the expectation is taken over the wage offer distribution $F(w|t)$ at t. Notice the similarity with Equation (4) above. The return of the asset $R(t)$ in a small interval around t equals the sum of the appreciation of the asset in this interval, the instantaneous utility flow in this interval, and the expected excess value of finding a job in this interval. The optimal strategy can be characterized by a reservation wage function $\phi(t)$ that gives the reservation wage at time t. Using the fact that $\phi(t) = \rho R(t)$, it follows that

$$\frac{\mathrm{d}\phi(t)}{\mathrm{d}t} = \rho\,\phi(t) - \rho\, b(t) - \lambda(t) \int_{\phi(t)}^{\infty} (w - \phi(t))\, \mathrm{d}F(w|t).$$

This differential equation has a unique solution for $\phi(t)$, given the boundary condition that follows from the assumption that the model is stationary for all sufficiently high t.

The hazard function $\theta(t)$ now equals

$$\theta(t) = \lambda(t)\, \overline{F}(\phi(t)|t).$$

In general, $\theta(t)$ varies with t. The distribution function for the duration of unemployment subsequently follows from Equation (1).

For examples of structural empirical analyses of nonstationary models with anticipation, see Wolpin (1987), Van den Berg (1990a), Engberg (1991) and Garcia-Perez (1998).

[7] Some special cases of this model have been examined earlier; see e.g., Mortensen (1986).

3.2. Repeated-search model

Models of repeated search allow the economic agent to search further for better matches after a match has been formed. The best-known model of repeated search is the so-called on-the-job search model which aims to describe the behavior of employed individuals who search for a better job [see Mortensen (1986) for an overview]. In the basic on-the-job search model, a job is characterized by its wage w which is taken to be constant within a job. For a working individual, the search environment is specified in exactly the same way as we did in Subsection 3.1.1 for an unemployed individual. In particular, we assume the model to be stationary. The optimal strategy is constant during a job spell, and the the expected present value of search $R(w)$ when following the optimal strategy in a job with wage w satisfies

$$\rho R(w) = w + \lambda \mathrm{E}_{w^*} \max\{0, R(w^*) - R(w)\},$$

where the expectation is taken with respect to the distribution F of wage offers w^*. Clearly, the optimal strategy is such that one accepts a job if and only if the offered wage w^* exceeds the current wage w, so it suffices to compare instantaneous income flows (i.e., the optimal strategy is “myopic”), and the reservation wage simply equals the current wage.

For a given current wage w, the hazard of the job duration distribution (or exit rate out of the present job) equals

$$\theta = \lambda \overline{F}(w).$$

As a result, the duration of a job with a wage w has an exponential distribution with this parameter θ. Note that models of repeated search are informative on the joint distribution of consecutive job durations.

If, during employment, exogenous separations occur at a rate δ, then this does not affect the optimal strategy. The exit rate out of the present job then equals $\lambda\overline{F}(w) + \delta$. See Flinn (1996) for an example of structural estimation of this model with job duration data[8].

Burgess (1989) introduces a rather manageable type of nonstationarity in this model. The individual’s search environment (i.e., λ and F) is subject to shocks that are not job-specific but rather such that they act similarly on all employed workers. The shocks may be anticipated or unanticipated. It is intuitively obvious that this nonstationarity does not change the optimal strategy: it remains optimal to accept another job if and

[8] The empirical analysis of so-called equilibrium search models, which endogenize the wage offer distribution F, often involves the joint estimation of the distributions of unemployment durations and job durations. See e.g., Van den Berg and Ridder (1998), Bontemps, Robin and Van den Berg (2000) and Bowlus, Kiefer and Neumann (2001).

only if its wage exceeds the current wage. We thus obtain for the job-to-job transition rate,

$$\theta(t) = \lambda(t)\,\overline{F}(w|t).$$

Throughout the remainder of the chapter, it is important to keep in mind that empirical duration analysis is ultimately interested in structural parameters that represent determinants of individual behavior. This is also true for empirical analysis in which reduced-form models are estimated that are not explicitly specified as a theoretical model. In the sequel we return to this issue.

4. The Mixed Proportional Hazard model

4.1. Definition

For the sake of convenience, we use the term "individual" to denote the subject that experiences certain spells in a given state. We consider the population of individuals that consists of the inflow into this given state. This can be the inflow at a given point of time, or the inflow at any time. We assume that, for a given individual in this population, the subsequent duration T is an absolutely continuous and positive random (duration) variable. The distribution of T (or, equivalently, the hazard function) may vary across individuals. We assume that all individual variation in the hazard function can be characterized by a finite-dimensional vector of observed explanatory variables (or "covariates", or "regressors") x and an unobserved heterogeneity term v. The latter term can be interpreted as a function of unobserved explanatory variables[9]. In this subsection we assume that x is time-invariant, and consequently we define the Mixed Proportional Hazard model as a model with time-invariant explanatory variables. In the next subsection we introduce time-varying explanatory variables.

For an individual with explanatory variables x and unobserved heterogeneity v, the hazard function of the random variable T evaluated at the duration t is denoted by $\theta(t|x, v)$. This notation highlights the fact that we condition on x and v. The standard MPH model is now defined by

Definition 1. *Standard MPH model*: There are functions ψ and θ_0 such that for every t and every x and v there holds that

$$\theta(t|x, v) = \psi(t) \cdot \theta_0(x) \cdot v. \tag{5}$$

This model was developed by Lancaster (1979), which includes an empirical application to unemployment duration data, and by Vaupel, Manton and Stallard

[9] Lancaster (1990) shows that v to some extent may also represent measurement errors in T and x.

(1979)[10]. The function $\psi(t)$ is called the "baseline hazard" since it gives the shape of the hazard function for any given individual. Only the *level* of the hazard function is allowed to differ across individuals. The term $\theta_0(x)$ is called the "systematic part" of the hazard. In applied work, it is common to specify

$$\theta_0(x) = \exp(x'\beta), \tag{6}$$

so that $\theta(t|x, \upsilon)$ is multiplicative in all separate elements of x.

For convenience, we make a number of regularity assumptions on the determinants of the model.

Assumption 1: *The vector x is k-dimensional with $1 \leqslant k < \infty$. The function $\theta_0(x) : \mathcal{X} \subset \mathbb{R}^k$ is positive for every $x \in \mathcal{X}$.*

Assumption 2: *The function $\psi(t)$ is positive and continuous on $[0, \infty)$, except that $\lim_{t \downarrow 0} \psi(t)$ may be infinite. For every $t \geqslant 0$ there holds that $\int_0^t \psi(\tau)\, \mathrm{d}\tau < \infty$, while $\lim_{t \to \infty} \int_0^t \psi(\tau)\, \mathrm{d}\tau = \infty$.*

Assumption 3: *The distribution G of υ in the inflow satisfies* $\Pr(0 < \upsilon < \infty) = 1$.

Assumption 4: *The individual value of υ is time-invariant.*

It should be stressed that, for virtually all of the results in the chapter, these conditions are stronger than needed. This is particularly true for Assumption 2. It is often sufficient that $\psi(t)$ is integrable, and sometimes it is sufficient that $\int_0^t \psi(\tau)\mathrm{d}\tau < \infty$ only on some interval. For expositional reasons, we do not deal with this. On the other hand, for identification, additional assumptions are needed (see Section 5). We do not list those here because it is interesting to contrast alternative assumptions in the light of identifiability issues.

It is useful to examine the special case in which there is no unobserved heterogeneity ($\upsilon \equiv 1$). In that case the model is called a Proportional Hazard (PH) model [this model was developed by Cox (1972) and predates the MPH model]. The PH model specification is regarded to be simple and yet sufficiently rich to capture many data properties. The popularity of the PH model in reduced-form duration analysis is comparable to the popularity of the linear regression model in reduced-form regression analysis. Note that the general regression-type expression for the integrated hazard function (see Equation 3) reduces to

$$\log \int_0^t \psi(u)\, \mathrm{d}u = -x'\beta + \varepsilon, \tag{7}$$

for the PH model, where we substituted Equation (6), and ε has an EV1 distribution. It should again be stressed that ε represents the purely random variation in the duration

[10] Nickell (1979) contains the first estimation of a discrete-time MPH-type model.

outcome – it does not capture unobserved individual characteristics. In comparison to a linear regression model (say $\log t = x'\beta + \varepsilon$, with ε having an unknown distribution with mean zero), the left-hand side of Equation (7) has a more general specification, since it involves an unknown transformation of the duration variable, whereas the right-hand side has a more restrictive specification, since the distribution of the error term is completely specified. Thus, the PH model and the regression model are not nested, and they derive their flexibility from different sources.

The β parameters in the linear regression model are estimated consistently by OLS under a wide range of distributions of ε. Similarly, the β parameters in the PH model are estimated consistently by Partial Likelihood under a wide range of specifications of the baseline hazard $\psi(t)$. More precisely, the β parameters are estimated consistently by maximization of a partial likelihood function that does not depend on the baseline hazard function, which can be estimated nonparametrically in a second stage [see Lancaster (1990) for details]. This is arguably one of the great advantages of the PH model, but it does not carry over to the MPH model in general.

For the MPH model, Equation (3) reduces to

$$\log \int_0^t \psi(u)\,\mathrm{d}u = -x'\beta - \log \upsilon + \varepsilon, \tag{8}$$

where again we substituted Equation (6), and where again ε has an EV1 distribution. The equation states that the log integrated baseline hazard function given x has the same distribution as the distribution of a random variable that is the sum of an EV1 random variable and another random variable (namely $-x'\beta - \log \upsilon$ given x). Since we have not made an assumption on the distribution of υ, it is clear that specification (8) is much more general than Equation (7). Now we have a flexible specification for both the transformation of t and the distribution of the error term. However, the latter distribution cannot be just any distribution. For example, it cannot be a normal distribution, because the sum of an EV1 random variable and another random variable cannot have a normal distribution [see Ridder (1990)]. It turns out that the MPH model is actually identified under an assumption on the tail of the distribution of υ (see Section 5).

We end this subsection by mentioning some other reduced-form duration models. Consider the following model,

$$\log z(t) = -x'\beta + \epsilon, \tag{9}$$

with $z(t)$ positive and increasing in t. This reduces to the MPH model if the "error term" ϵ is distributed as the sum of an EV1 random variable and another random variable. If no assumption is made on the distribution of ϵ then Equation (9) is called a "transformation model" [see Horowitz (1996)]. If it is subsequently imposed that $z(t) = t$ then we obtain the Accelerated Failure Time (AFT) model,

$$\log t = -x'\beta + \epsilon.$$

For future reference it is useful to note that in the AFT model the survivor function can be written as

$$\overline{F}(t|x) = \exp\left(-\Psi\left(t\cdot \mathrm{e}^{x'\beta}\right)\right), \tag{10}$$

where Ψ is the integrated hazard function of the random variable $\exp(\epsilon)$. Clearly, the individual characteristics act on the duration distribution by transforming the time scale from t to $t\exp(x'\beta)$. This may be an accurate description of the actual variation in the lifetime distributions of complex self-evolving organisms or mechanisms. Because of the one-to-one relation between a distribution and its hazard function, the AFT specification can be translated into a specification of the hazard function of $t|x$. Obviously, the latter need not be an MPH specification. Note that in the transformation model and the AFT model, the hazard does not serve as the focal point of model specification. This has strongly limited the use of these models in social science duration analyses. We return to this in Subsection 5.6.

4.2. Time-varying explanatory variables

In practice, explanatory variables are often time-varying, and there are often good reasons to assume that the hazard function is affected by the current value of the explanatory variable (instead of, e.g., its value at the beginning of the spell). In this subsection we discuss the incorporation of such explanatory variables in the PH model and (at the end of the subsection) the MPH model. Given that the chapter avoids measure theory, the exposition in this subsection is restricted to be rather informal, and we refer the reader to the references below for more rigorous analyses.

At first sight it may seem that time-varying explanatory variables can be incorporated in the PH model by replacing x by $x(t)$,

$$\lim_{\mathrm{d}t\downarrow 0}\frac{\Pr(T\in[t,t+\mathrm{d}t)|T\geqslant t,\{x(u)\}_0^t)}{\mathrm{d}t} = \psi(t)\cdot\theta_0(x(t)), \tag{11}$$

where $\{x(u)\}_0^t$ denotes the time path of x up to t, and where $\theta_0(x(t)) = \exp(x(t)'\beta)$, possibly. However, there are some caveats here. First, the values of the explanatory variables at t may in some sense be endogenous. The subject under study may have inside information at t on the future realization of the random variable T, and this information may affect the values of his observed explanatory variables at t and his hazard rate at t. It may then be erroneously concluded that the observed explanatory variables have a causal effect on the duration. Consider an unemployed individual who knows that he will start to work in a job at a given future date and may for that reason decide not to enrol in a training program at t. If this is ignored in the empirical analysis then the effect of the number $x(t)$ of completed training programs at t on the exit rate out of unemployment at t may be under-estimated. A second caveat concerns the fact that $x(t)$ could cause the duration distribution to be discontinuous at certain durations. This would complicate the statistical and empirical analysis.

To proceed, assume that the time-varying explanatory variables constitute a stochastic process $X = \{X(t) : t \geqslant 0\}$. Without loss of generality we take $X(t)$ to represent *all* explanatory variables for the hazard rate at t. Note that we may trivially include time-invariant or fully deterministic explanatory variables in X, and recall that for the time being we assume that all heterogeneity is observed. Kalbfleisch and Prentice (1980) develop a classification of duration models with time-varying covariates, in order to describe classes for which standard econometric procedures can be applied. This classification is rather vague and not exhaustive [Heckman and Taber (1994)]. Fortunately, the recent mathematical-statistical literature on counting processes and martingales has allowed a breakthrough on these issues. The counting process approach assumes that the durations, the values of the time-varying explanatory variables, and the observational plan, are all outcomes of stochastic processes [as such, it allows for quite general censoring schemes; see Fleming and Harrington (1991), Andersen and Borgan (1985) and Andersen et al. (1993) for excellent surveys, and Ridder and Tunalı (1999) for an exposition which also avoids measure theory and includes an econometric application]. The approach focuses on a PH model framework in which X has the property that:

- X is a predictable process.

Here, predictability basically means that the values of all explanatory variables for the hazard at t must be known (and observable to the researcher) just before t. In other words, the values of the variables which capture all individual variation in the hazard rate at t must be known and observable at t^{-}. In other words, the values of the explanatory variables at t are influenced only by events that have occurred up to time t, and these events are observable. The information on the values *at* time t does not help in predicting a transition at t. Note that predictability does *not* mean that the whole future realization of X can be predicted at some point in time. Below we give some examples. Ridder and Tunalı (1999) argue that the concept of predictability is basically the same as the concept of weak exogeneity in time series analysis (and is thus weaker than the concept of strong exogeneity). In addition to predictability, we need a technical assumption which basically ensures that the realized outcomes of $X(t)$ and $\theta_0(X(t))$ are bounded. Fleming and Harrington (1991) contains a more precise exposition with explicit use of measure theory. The counting process approach has been very successful in the derivation of (asymptotic) properties of estimators and test statistics for general settings, including generalizations of the commonly used estimators and test statistics in duration analysis (see the references above).

Now consider the stochastic process $\Pr(T \leqslant t | \{X(u)\}_0^t)$, which is a process given the evolution of X up to t, as a function of t. Assume that this process is absolutely continuous. Often, sufficient for this (in addition to the predictability of X) is, basically, that T does not have a strictly positive probability of occurrence at t, given X up to t. Given absolute continuity, the counting process model can be expressed as a model of hazard functions. Conversely, a PH model of hazard functions, with X having the above properties, and with absolute continuity of the above process, can be thought of as being generated by a PH counting process model [Fleming and Harrington (1991), Arjas (1989)]. It should be noted that these results have been derived for models with

$$\theta_0(X(t)) = \exp(X(t)'\beta),$$

and certain other specifications of θ_0 [see Andersen and Borgan (1985)].

The results imply that if we start off with a PH-type model of a hazard function, and X has the properties above, then we can perform valid econometric inference using standard methods, on the basis of specification (11) for the hazard rate. This is, in a nutshell, why predictability of the time-varying explanatory variable is an extremely useful property. Given predictability, we may apply the standard tools of duration analysis[11].

It is useful to examine the predictability for some special cases for X. First, if X is time-invariant then it is obvious that it is predictable. Now suppose its path is fully known in advance. For example, the unemployment benefits level as a function of the elapsed unemployment duration may be determined at the date of inflow into unemployment, by the institutional setting. Clearly, X is then predictable as well. If X is stochastic then somewhat loosely one may state that if the current value of X only depends on past and outside random variation then X is predictable [Andersen and Borgan (1985)]. Now consider the case in which the individual has inside information on future realizations of X. For example, an unemployed individual may expect a baby or may expect participation in a training program at a future date. This information may be used as input in the individual's decision problem and as a result may affect the current hazard rate. If this information is not known to the analyst then X is not predictable. The same is true if the individual anticipates the realization of T and if this affects the current hazard. Note that it is intuitively plausible that, in these cases, standard inference may lead to inconsistent estimates. These cases include so-called instantaneous feedback effects: predictability is not satisfied if X jumps in an unexpected way at t. This does not mean that jumps in regressor values are not allowed at all if one demands predictability. Suppose that one wants to model that an individual's hazard rate increases by a certain amount immediately after the realization of another duration variable τ which is otherwise independently distributed from the duration of interest and from other time-varying covariates. This can be captured by a time-varying regressor $\mathrm{I}(t > \tau)$, which is predictable.

Now consider the case where a time-invariant explanatory variable is unobserved (i.e., consider MPH models). If we condition on the unobserved heterogeneity value υ and do as if υ is observed, then the above analysis remains valid. If υ is treated as unobserved then υ is not predictable. As we shall see in Section 5, ignoring

[11] Note that in case of stochastic explanatory variables it does not make sense to talk about "the" probability distribution of T.

the unobserved heterogeneity in empirical inference generally leads to inconsistent inference. In this case, the standard solution is to jointly model the hazard function and the distribution of v, and to integrate v out of the likelihood.

We end this subsection by making a few comments. First, time-varying explanatory variables may play a very different role in *other* reduced-form duration models, such as the AFT model. This reflects the fact that such models do not take the hazard function as the point of departure for the model specification[12]. Secondly, as noted above, the counting process approach allows for quite general censoring schemes; in fact, what is needed is that the observational plan is a predictable process. Thirdly, in the remainder of the chapter, the focus is mostly on models without time-varying explanatory variables. The motivation for this is basically the same as the one (implicitly) adopted in most of the methodological literature on duration models, namely that the analysis of these models is relatively manageable and that the results create a good starting point for future analysis of more general models. Below, whenever we encounter time-varying explanatory variables, we tacitly assume that the conditions that ensure valid inference with standard methods are satisfied.

4.3. Theoretical justification

As mentioned above, the MPH model and its special cases are often regarded to be useful reduced-form models for duration analysis. The resulting estimates are generally interpreted with the help of some economic theory. However, the MPH model specification is not derived from economic theory, and it remains to be seen whether the MPH specification is actually able to capture important theoretical relations, and, conversely, whether the MPH specification can be generated by theory.

The main assumption underlying the MPH model is that the three determinants of the hazard act multiplicatively on the hazard. This implies that if the elapsed duration has a positive effect on the hazard, then this effect is stronger for individuals with characteristics that also have a positive effect on the hazard. Of course, the distinction between two of the three determinants (the observed and unobserved explanatory variables) is only relevant from an empirical point of view. If the researcher could observe all determinants without measurement error, then the unobserved heterogeneity term can be omitted. Within a theoretical framework it is irrelevant whether a certain background variable can be observed by the researcher or not. This means that from a theoretical point of view, the most important assumption

[12] For example, consider the formulation (10) of the AFT model. Typically, time-varying explanatory variables are included in this model by way of

$$\overline{F}(t|\{X(u)\}_0^t) = \exp\left(-\Psi\left(\int_0^t \exp(X(u)'\beta)\,\mathrm{d}u\right)\right).$$

In that case, the hazard rate at t depends on the whole history $\{X(u)\}_0^t$ of X.

of the MPH model is that the elapsed duration and the explanatory variables act multiplicatively on the hazard.

In economics, this assumption is often hard to justify. We illustrate this by examining the economic theories discussed in Section 3[13]. First consider the job search model of Subsection 3.1.2. We allow all structural determinants to differ across individuals, and this is captured by time-invariant explanatory variables x. We assume that the analyst observes x (and the duration t) but does not directly observe how the structural determinants, the optimal strategy, or the acceptance probability change with t. If such changes would be directly observed then obviously it would make sense to include them as time-varying explanatory variables. We return to time-varying explanatory variables towards the end of the subsection.

From Subsection 3.1.2 we obtain the following system of equations, in obvious notation,

$$\phi(t,x) = b(t,x) + \frac{\lambda(t,x)}{\rho(t,x)} \int_{\phi(t,x)}^{\infty} \overline{F}(w|t,x)\,\mathrm{d}w, \qquad \theta(t,x) = \lambda(t,x)\overline{F}(\phi(t,x)|t,x).$$

Intuitively, the main reason for why it is difficult to obtain a multiplicative structure for $\theta(t,x)$ is that in general $\overline{F}(\phi(t,x)|t,x)$ is not multiplicative in ϕ, which in turn depends on "everything in the model" in a non-multiplicative fashion. Below are a few special cases where the resulting $\theta(t,x)$ *is* proportional in t and x. Note that these assume that changes in the structural determinants are unanticipated.

Example 1. Let F be a Pareto distribution,

$$\begin{aligned}&\text{Family of Pareto distributions:}\\ &\overline{F}(w) = (w_0/w)^{\nu} \quad \text{for all } w > w_0, \text{ with } w_0, \nu > 0,\end{aligned} \tag{12}$$

where we actually assume $\nu > 1$ to ensure that the optimal strategy exists, and where the parameters w_0 and ν of F may depend on t and x. Let in addition $b \equiv 0$. Then

$$\theta(t,x) = \rho(t,x)(\nu(t,x) - 1).$$

Let the discount rate ρ vary with x but not with t, and let the shape parameter ν vary with t but not with x (for example, long-term unemployed workers receive on average lower wage offers). Then the hazard is proportional in t and x. Of course, the same result applies if ρ only varies with t and ν only with x. Also, if ν is a fixed constant and ρ is proportional in t and x, then the hazard is proportional as well. Note that the assumption $b \equiv 0$ is very strong[14].

[13] The problem is more general, though.

[14] In general, if one is prepared to adopt a linearized specification for the reservation wage $\phi(t,x)$ as a function of its determinants, and if F has a Pareto distribution or an exponential distribution, then it is less difficult to obtain a multiplicative specification for $\theta(t,x)$ [see Lancaster (1985a)].

Example 2. Let $\rho = \infty$, so that workers do not care about the future. Then $\phi \equiv b$, and

$$\theta(t,x) = \lambda(t,x)\overline{F}(b(t,x)|t,x).$$

If $\lambda(t,x)$ varies with t (e.g., because the long-term unemployed are stigmatized) but not with x, and F and b vary with x but not with t, then the hazard is proportional in t and x. Alternatively, if F and b do not depend on either t or x and λ is proportional in t and x, then the hazard is proportional as well.

Example 3. Let the structural determinants be such that ϕ is always smaller than the lowest wage in the market (e.g., benefits are so low that the reservation wage is below the mandatory minimum wage). Then $\overline{F}(\phi) = 1$ always, and

$$\theta(t,x) = \lambda(t,x),$$

so, if λ is proportional in t and x, then the hazard is proportional as well.

Example 4. This case is based on Yoon (1985), which is one of the very few studies to date on the theoretical justification of the PH model. He examines a model where jobs have a fixed and common tenure T^*, after which the individual dies[15]. The variable b is assumed to equal benefits minus search costs, and the model requires that the net value of b is negative. There is no discounting of the future (so the limiting case $\rho \downarrow 0$ is considered). It is straightforward to show that $\phi(t)$ then follows from

$$-b(t,x) = \lambda(t,x)T^* \int_{\phi(t,x)}^{\infty} \overline{F}(w|t,x)\,\mathrm{d}w.$$

Let F be a Pareto distribution (see Equation 12) with a fixed parameter $\nu > 1$ and a parameter $w_0(t,x)$. It follows that

$$\theta(t,x) = [\lambda(t,x)]^{\frac{-1}{\nu-1}}\,[w_0(t,x)]^{\frac{-\nu}{\nu-1}}\,\left[\frac{-b(t,x)(\nu-1)}{T^*}\right]^{\frac{\nu}{\nu-1}}.$$

Obviously, there are many ways to obtain a PH specification from this.

Now consider anticipated changes in the structural determinants, i.e., consider the nonstationary job search model of Subsection 3.1.3. In particular, for ease of exposition, consider a special case where the only change concerns a drop in b at a

[15] Job separations leading to unemployment rather than death or permanent retirement are hard to reconcile with unanticipated duration dependence of the structural determinants, because of the repetitive nature of unemployment.

duration τ (from b_1 to b_2). There still holds that $\theta(t,x) = \lambda(t,x)\overline{F}(\phi(t,x)|t,x)$. However, now the reservation wage $\phi(t)$ for $t < \tau$ depends on b_1 and b_2 as well as on $\tau - t$. The smaller the remaining time interval $\tau - t$ until the drop in b, the more important the future benefits level b_2 is for the current present value. As shown by Van den Berg (1990a, 1995), there are two reasons for this. First, the discounting of the future means that the far future carries less weight than the near future. Second, there is a probability that the individual leaves unemployment before τ, and this probability is lower if τ is in the near future. This probability depends on the hazard function itself, in between t and τ. As a result of all this, as the duration $t < \tau$ proceeds, the effect on the hazard of b_1 diminishes, and the effect of b_2 increases (with a magnitude that depends on all structural determinants). After τ, the hazard does not depend on b_1 anymore. It seems to be impossible to justify a PH specification with such a theoretical model, except for the following limiting case.

Example 5. Let $\rho \to \infty$ in the nonstationary job search model, so workers do not care about the future. In that case, even though an individual does have information on future changes, this does not affect his optimal strategy, and the exit rate out of unemployment is the same as in Example 2.

Finally, consider the nonstationary on-the-job search model of Subsection 3.2, and, in particular, the job-to-job transition rate (which will be our hazard rate). Note that there is no "feedback" from the structural determinants to the value of the reservation wage w. There holds that $\theta(t,x) = \lambda(t,x)\overline{F}(w|t,x)$, where x may include w, and the following result emerges.

Example 6. Let F be time-invariant in the nonstationary on-the-job search model. Then

$$\theta(t,x) = \lambda(t,x)\overline{F}(w|x),$$

which supports a PH specification if $\lambda(t,x)$ is multiplicative in t and x. If F has a Pareto distribution (see Equation 12), then its parameter w_0 is allowed to depend on t.[16, 17]

The main conclusions of this subsection are as follows. First, the proportionality restriction of the (M)PH model can in general not be justified on economic-theoretical grounds. Second, if the optimal strategy is myopic (e.g., because of repeated search, or

[16] The proportionality results in Examples 4 and 6 can also be generated with other families of wage offer distributions than the Pareto family. Notably, F can be exponentially distributed, so $\overline{F}(w) = \exp(-v(w - w_0))$ on $w > w_0$, with $v > 0$.

[17] Here, as in previous examples, if the job offer arrival rate depends on an optimally chosen search intensity, then the scope for multiplicative specifications is further reduced. This is because this search intensity is a second "channel" through which all structural determinants affect the hazard in a non-multiplicative fashion [see e.g., Mortensen (1986) for a theoretical analysis of such models].

because the discount rate is infinite), then this restriction often follows from economic theory.

Despite the first conclusion, the (M)PH model has become very popular in reduced-form duration analysis, in particular in labor economics. The popularity of a reduced-form model that does not nest many structural models distinguishes duration analysis from the reduced-form analysis of wage data with the linear regression model, since the linear specification has been justified extensively by human capital theory and traditional labor supply theory. Part of the attractiveness of the (M)PH model stems from the fact that it is difficult to think of a more parsimonious specification of the hazard that includes all the single major determinants. (Also, recall that the Partial Likelihood estimation method allows for estimation of the systematic hazard of the PH model without the need to parameterize or estimate the baseline hazard.) In practice, the empirical application at hand does not always dictate a natural theoretical framework, and sometimes the scope of the application does not warrant a full-blown theoretical or structural analysis. In such cases, the (M)PH model is a useful framework whose properties have been thoroughly studied in the literature.

Last but not least, the MPH framework can be extended to a certain extent to incorporate some features of the theory at hand. Notably, changes over t in the value of a variable x can be incorporated by the inclusion of time-varying covariates. For example, in the study of unemployment insurance benefits on exit out of unemployment, the effect of the remaining benefit entitlement can be included as a time-varying covariate [see e.g., Solon (1985)]. Also, if the data provide direct observations on how a structural determinant, the reservation wage, or the acceptance probability change over time, then these can be included as time-varying covariates. As an example, consider the models of Subsection 3.1, and suppose that $\phi(t,x)$ is fully observed and F is a time-invariant Pareto distribution which does not vary with x. Then $\theta(t,x) = \lambda(t,x)w_0^{\nu}[\phi(t,x)]^{-\nu}$, so if $\lambda(t,x)$ is multiplicative in t and x then this supports a PH specification with a time-varying covariate. As another example, consider the on-the-job model. One may observe business cycle indicators and use these as representations of $\lambda(t,x)$. Finally, changes in the effect over t of a variable x can be incorporated by the inclusion of interactions between t and x in the hazard[18].

These extensions lead to less transparent models, and some of the distinct advantages of the MPH model are lost this way (see Section 5). Moreover, it should be stressed that the insertion of some time-varying covariates or time-varying parameters into an MPH model more often than not does not lead to a specification that can be generated by a theoretical model. This is intuitively clear from the nonstationary model in which unemployment benefits decrease with the duration of unemployment.

As noted in Subsection 4.1, in applied work it is often assumed that each explanatory variable acts multiplicatively on the hazard rate (i.e., $\theta_0(x) = \exp(x'\beta)$). From the

[18] One may use a nonparametric estimation method for an unrestricted specification of the hazard rate $\theta(t,x)$, allowing for full interactions [see e.g., Dabrowska (1987)].

discussion above it is clear that economic theory often predicts that the different structural determinants do not act multiplicatively on the hazard. Thus, if each determinant is represented by different elements of x, then these elements interact with each other in the hazard. This can be incorporated to a certain extent in the MPH model, as inclusion of interaction terms for the different elements of x does not violate the (M)PH specification[19].

We end this section by noting that the economic justification of other popular reduced-form duration model specifications is at least as difficult as the justification of the (M)PH specification. This holds in particular for the Accelerated Failure Time model, in which the mean of $\log t$ is specified as a linear function of x, so $\log t = -x'\beta + \epsilon$, and also for the additive hazard model, in which $\theta(t|x)$ is specified as $\theta(t|x) = \psi(t) + \theta_0(x)$. These two types of reduced-form duration models enjoy popularity in biostatistics, where the relation between theory and application is less compelling than in econometrics. Discrete-time reduced-form duration model specifications are also difficult to justify; they often do not follow from the underlying economic models (like discrete-time search models or dynamic discrete-choice models).

5. Identification of the MPH model with single-spell data

5.1. Some implications of the MPH model specification

In this section we examine identification of the MPH model with unobserved heterogeneity[20], if the data provide i.i.d. drawings from the conditional distribution of $t|x$. In reality, the observations on t may be right-censored (i.e., for some observations it is only known that t exceeds a certain value) or interval-censored (e.g., if durations are grouped into intervals), or the sampling design may be non-random. Heckman and Singer (1984a), Ridder (1984) and Lancaster (1990) contain extensive examinations of the implied duration distributions in other sampling designs. Situations in which the data provide multiple durations for the same individual are discussed in subsequent sections.

Throughout the section we make the following model assumption,

Assumption 5. Independence of observed and unobserved explanatory variables: *In the inflow, v is independent of x.*

[19] As an example, job search theory predicts that the elasticity of the exit rate out of unemployment with respect to unemployment benefits depends on the level of the benefits. This can be captured to some extent in a reduced-form analysis by including $(\log b)^2$ as an additional regressor [see Van den Berg (1990c) for details].

[20] Identification of the determinants of the PH model is trivial if it is known that the data are generated by a PH model.

Note that this assumption is stronger than the usual assumption in linear regression models that x and ε are uncorrelated or that they satisfy $\mathrm{E}(\varepsilon|x) = 0$.

It is useful to examine the distribution $F(t|x)$ of $t|x$ and derive the well-known result that the duration dependence of the hazard function $\theta(t|x)$ of this distribution is more negative than the duration dependence of the hazard function $\theta(t|x, v)$ [Lancaster (1979) was the first point out these results; see also the survey in Lancaster (1990) and Heckman and Singer (1984a), who consider a generalization of the MPH framework].

By definition, we have

$$F(t|x) = \int_0^\infty F(t|x, v)\, \mathrm{d}G(v), \tag{13}$$

where G is the cumulative distribution function of v in the inflow into the state of interest, and where $F(t|x, v)$ has the associated hazard function $\theta(t|x, v)$. Consequently, $\theta(t|x)$, which by definition equals $f(t|x)/\overline{F}(t|x)$, can be written as

$$\theta(t|x) = \frac{\int_0^\infty \theta(t|x, v)\, \overline{F}(t|x, v)\, \mathrm{d}G(v)}{\overline{F}(t|x)}. \tag{14}$$

By Bayes' Theorem, we have for every t that

$$\mathrm{d}G(v|T > t, x) = \frac{\overline{F}(t|x, v)\, \mathrm{d}G(v)}{\overline{F}(t|x)}. \tag{15}$$

(Note that here we use T to denote a random variable.) In general, therefore, the distribution of $v|T > t, x$ depends on x for all $t > 0$, even though it does not for $t = 0$. The composition of the sample of survivors (as captured by the distribution of v) changes as time proceeds, in a way that that depends on t and x. This is an important aspect of the dynamic self-selection that occurs if one examines subsamples of individuals with higher and higher durations.

Substitution of Equation (15) into Equation (14) yields $\theta(t|x) = \mathrm{E}_{v|T>t,x}(\theta(t|x, v))$. Therefore,

$$\theta(t|x) = \psi(t) \cdot \theta_0(x) \cdot \mathrm{E}(v|T > t, x). \tag{16}$$

Let us denote the integrated baseline hazard at t as $z(t)$,

$$z(t) = \int_0^t \psi(\tau)\, \mathrm{d}\tau.$$

Of course, $-\log \overline{F}(t|x, v)$ equals $v \cdot \theta_0(x) \cdot z(t)$. By substituting this into Equations (13) and (15) it follows that we can write

$$\mathrm{E}(v|T > t, x) = \frac{\int_0^\infty v \cdot \mathrm{e}^{-v \cdot \theta_0(x) \cdot z(t)}\, \mathrm{d}G(v)}{\int_0^\infty \mathrm{e}^{-v \cdot \theta_0(x) \cdot z(t)}\, \mathrm{d}G(v)}. \tag{17}$$

It is useful to rewrite $\theta(t|x)$ in some different ways. First, note that the denominator on the right-hand side of Equation (17) (which equals $\overline{F}(t|x)$) is nothing but the Laplace transform $\mathcal{L}$ of the distribution of υ, evaluated at $\theta_0(x) \cdot z(t)$,

$$\mathcal{L}(s) = \int_0^\infty \mathrm{e}^{-s \cdot \upsilon}\, \mathrm{d}G(\upsilon). \tag{18}$$

Consequently, the numerator in Equation (17) is nothing but minus the derivative of $\mathcal{L}$ evaluated at $\theta_0(x) \cdot z(t)$. This means that we can rewrite Equation (16) as follows,

$$\theta(t|x) = \psi(t) \cdot \theta_0(x) \cdot \frac{-\mathcal{L}'(\theta_0(x) \cdot z(t))}{\mathcal{L}(\theta_0(x) \cdot z(t))}. \tag{19}$$

So all derivatives of this with respect to x and/or t depend on G only by way of (derivatives of) the Laplace transform of G, evaluated at $\theta_0(x) z(t)$. Equivalently, all derivatives of $\theta(t|x)$ with respect to x and/or t depend on G by way of moments of $\upsilon|T > t, x$. Specifically,

$$\frac{\mathrm{d} \log \theta(t|x)}{\mathrm{d}t} = \frac{\psi'(t)}{\psi(t)} - \frac{\mathrm{Var}(\upsilon|T > t, x)}{\mathrm{E}(\upsilon|T > t, x)} \cdot \psi(t)\, \theta_0(x).$$

Clearly, because of the presence of unobserved heterogeneity (i.e., $\mathrm{Var}(\upsilon) > 0$, which under regularity conditions implies that $\mathrm{Var}(\upsilon|T > t, x) > 0$), the duration dependence in the observed (or "aggregate") hazard function $\theta(t|x)$ is more negative than otherwise. This is because in case of unobserved heterogeneity, the individuals with the highest values of υ (and thus the highest hazards) on average leave the state quickest, so that the individuals who are still in this state at high durations tend to have lower values of υ and thus lower hazards. This phenomenon has been called "weeding out" or "sorting". It occurs in duration models with unobserved heterogeneity in general, and so is not restricted to the MPH model. The model thus allows for two competing explanations for observed negative duration dependence. If one ignores the presence of unobserved heterogeneity (i.e., if one adopts a PH model whereas the data are generated by an MPH model with $\mathrm{Var}(\upsilon) > 0$), then the estimated duration dependence will be too negative. This result has spurred the literature on the identification of duration models with unobserved heterogeneity.

Unobserved heterogeneity has a similar effect on the derivative of $\log \theta(t|x)$ with respect to x,

$$\frac{\mathrm{d} \log \theta(t|x)}{\mathrm{d}x} = \frac{\theta_0'(x)}{\theta_0(x)} - \frac{\mathrm{Var}(\upsilon|T > t, x)}{\mathrm{E}(\upsilon|T > t, x)} \cdot z(t)\, \theta_0'(x). \tag{20}$$

Note that in the case $\theta_0(x) = \exp(x'\beta)$, the first term on the right-hand side reduces to β, and $\theta_0'(x)$ in the second term reduces to $\theta_0(x)\beta$. Because of the presence of unobserved heterogeneity, the semi-elasticity of the observed hazard function $\theta(t|x)$ with respect to x is closer to zero than otherwise. This can be understood as follows. Within the group of individuals with a high value of $\theta_0(x)$, the weeding out induced by unobserved heterogeneity goes much faster than within the group of individuals with a low value of $\theta_0(x)$. This is a consequence of the multiplicative specification of $\theta(t|x, v)$: a high $\theta_0(x)$ and a high v reinforce each other in producing a very high hazard. As a result, at a given duration $t > 0$, the sample of survivors with high $\theta_0(x)$ has on average lower values of v than the sample of survivors with low $\theta_0(x)$. This causes the observed average difference between the hazards of the survivors of these groups to be smaller than the true average difference between the two groups. It is important to stress that this does not automatically imply that, if one ignores the presence of unobserved heterogeneity while estimating the model with Maximum Likelihood, that then the effect of x on the individual hazard is under-estimated. This is basically because β has one more element than x, and the ML estimates of β are jointly determined. We return to this in Subsection 5.6.

Note that if $\mathrm{E}(v) < \infty$ and $\psi(0) < \infty$ then for $t = 0$ the right-hand side of Equation (16) reduces to $\psi(0)\,\theta_0(x)\mathrm{E}(v)$, and the function $\theta_0(x)$ is then identified from data on $\theta(0|x)$. This makes sense, as at $t = 0$ there is not yet any self-selection due to weeding out. Before we proceed with the identification of the full model (i.e., of the functions ψ, θ_0 and G), it is useful to introduce the function $h(s)$, defined as $-\mathcal{L}'(s)/\mathcal{L}(s)$ (see Equation 18). Equation (19) can now be rewritten as

$$\theta(t|x) = \psi(t) \cdot \theta_0(x) \cdot h(z(t)\,\theta_0(x)). \tag{21}$$

This equation will be useful in Subsection 5.3 and further.

5.2. *Identification results*

There is a substantial literature on the identification of the MPH model[21]. It is important to stress that no parametric functional form assumptions are made on the underlying functions θ_0, ψ and G, so the literature is concerned with *nonparametric* identification. In general, it is assumed that the data provide the distribution function $F(t|x)$ for all t and x.

It is useful to define identifiability as a property of the mapping from the determinants ψ, θ_0 and G, given their domain, to the data (as summarized in $F(t|x)$ for all t and x). Consider a given set of assumptions on the three determinants (like the restriction that their function values must be nonnegative; below we examine various sets of assumptions). These characterize the domain of the mapping. The MPH specification then defines the unique mapping from the domain to the data. The

[21] Heckman (1991) provides an overview in which the MPH model is embedded in a more general class of models. Heckman and Taber (1994) list identification proofs for MPH models, non-MPH models, and more tightly specified MPH models without covariates.

model is identified if the mapping has an inverse, i.e., if for given data[22] there is a unique set of functions ψ, θ_0 and G in the domain that is able to generate these data[23].

Now let us consider the assumptions that are made on the determinants. These include the regularity Assumptions 1–4, and Assumption 5 on the independence of x and v. In addition, we list the following assumptions which will play a role in the remainder of the chapter:

Assumption 6. Variation in observed explanatory variables: *The set $\mathcal{X}$ of possible values of x contains at least two values, and $\theta_0(x)$ is not constant on $\mathcal{X}$.*

Assumption 6b. Variation in observed explanatory variables: *There is an element x^a of the vector x with the property that the set $\mathcal{X}^a$ of its possible values contains a non-empty open interval. For given values of the other elements of x, the value of x^a varies over this interval. Moreover, $\theta_0(x)$ as a function of x^a is differentiable and not constant on this interval.*

Assumption 7. Normalizations: *For some a priori chosen t_0 and x_0, there holds that $\int_0^{t_0} \psi(\tau)\,\mathrm{d}\tau = 1$ and $\theta_0(x_0) = 1$.*

Assumption 8. Tail of the unobserved heterogeneity distribution: $\mathrm{E}(v) < \infty$.

Assumption 8b. Tail of the unobserved heterogeneity distribution: *The random variable v is continuous, and the probability density function $g(v)$ of v has the property that*

$$\lim_{v \to \infty} \frac{g(v)}{v^{-1-\epsilon} S(v)} = 1, \tag{22}$$

where $\epsilon \in (0, 1)$ is specified in advance, and where $S(v)$ is a slowly varying function[24], i.e., S has the property that, for every $v > 0$,

$$\lim_{u \to \infty} S(uv)/S(u) = 1.$$

For Assumption 6, a single dummy variable x suffices, provided that it has an effect on the hazard function. In that case $\theta_0(x)$ takes on only two values on $\mathcal{X}$. Note that we define θ_0 to be identified if its value is known for each $x \in \mathcal{X}$. In practice, one may start off with a parametric specification of $\theta_0(x)$ and require that all parameters

[22] Of course, these data must be in the image of the mapping.

[23] In fact, for technical reasons, the identification literature typically focuses on the model determinant z instead of its derivative ψ.

[24] See Feller (1971) for an exposition on such functions.

can be recovered from the set of all pairs $(x, \theta_0(x))$ with $x \in \mathcal{X}$. In the case where $\theta_0(x)$ is (log-)linear in $x'\beta$, this implies that the elements of x should not be perfectly collinear.

Assumption 7 concerns an innocuous normalization of two of the three terms in the hazard $\theta(t|x, v)$. Assumptions 8 and 8b require more discussion. Basically, under Assumption 8, the right-hand tail of G is not allowed to be too fat because otherwise $\mathrm{E}(v) = \infty$. Now consider Assumption 8b. It is important to stress that the a priori choice of ϵ determines the assumed class of heterogeneity distributions. Basically, the smaller ϵ, the fatter the tails. However, for any $\epsilon \in (0, 1)$, all heterogeneity distributions have $\mathrm{E}(v) = \infty$ [see Ridder (1990)]. This means that the right-hand tail of G is always fatter than under Assumption 8.

Elbers and Ridder (1982) were the first to prove the nonparametric identification of the MPH model, under Assumptions 1–8. Their identification proof is not constructive, i.e., the proof does not express the underlying functions θ_0, ψ and G directly in terms of observable quantities. Constructive identification proofs are attractive because they suggest a nonparametric estimation method. Melino and Sueyoshi (1990) provide a constructive proof for the case where Assumption 6 is tightened (to Assumption 6b, with the exception that $\theta_0(x)$ does not have to be differentiable). However, this proof is difficult to use as an inspiration for an attractive estimation strategy because it relies heavily on the observed duration density at $t = 0$, and x needs to be a continuous variable. Recently, Kortram et al. (1995) provided a constructive proof for the original case with only two possible values for $\theta_0(x)$. Lenstra and Van Rooij (1998) exploit this to construct a consistent nonparametric model estimator. They do not provide the asymptotic distribution of their estimator. Under somewhat stronger model assumptions than above, Horowitz (1999) constructs a nonparametric estimation method that does not follow an identification proof; rather, it exploits the similarity between the MPH model and the transformation model (see Subsection 4.1)[25,26]. He does provide the asymptotic distribution of his model estimator.

Heckman and Singer (1984b) also prove nonparametric identification of the MPH model. Their result turns out to be particularly interesting for the insights it generates into fundamental properties of the MPH model. Contrary to Elbers and Ridder (1982), they make Assumption 6b instead of the weaker Assumption 6, on the variation in x. More importantly, they make Assumption 8b instead Assumption 8 on the class of heterogeneity distributions. Assumption 8b rules out that v is degenerate. This means that the PH model as an underlying model is not included in the set of

[25] In fact, Horowitz (1999) assumes that $\theta_0(x) = \exp(x'\beta)$, and he accordingly calls the estimator a semiparametric estimator. It should be stressed that this estimator and other nonparametric and semiparametric estimators for the MPH model rely heavily on the shape of the empirical survivor function for $t \downarrow 0$. For a number of reasons, it is notoriously difficult to assess this shape. For example, extremely short durations are often under-reported in real-life data.

[26] Horowitz (1999) also provides a useful list of existing semiparametric estimation methods where parametric functional forms are assumed for either ψ or G.

MPH models considered by Heckman and Singer (1984b). This is a disadvantage if the PH model is regarded to be an interesting special case. This result should *not* be taken to mean that the MPH models considered by Heckman and Singer (1984b) are not able to generate a PH specification for the *observed* hazard $\theta(t|x)$. Consider the set of MPH models generated by a particular choice of ϵ in Equation (22), and assume that υ has a Positive Stable distribution. This family of distributions is most easily characterized by its Laplace transform.

Family of Positive Stable distributions:
$\mathcal{L}(s) = \exp(-s^{\alpha}) \quad$ with $\alpha \in (0, 1)$.

Note that $\lim_{s\downarrow 0}\mathcal{L}'(s) = -\infty$, so $\mathrm{E}(\upsilon) = \infty$[27]. Using results in Ridder (1990) and Feller (1971) it can be shown that in fact we have to take α exactly equal to ϵ in order to obtain a G that satisfies Equation (22). So, let υ have a Positive Stable distribution with parameter ϵ. Then, by Equation (19),

$$\theta(t|x) = \alpha\psi(t)[z(t)]^{\alpha-1}[\theta_0(x)]^{\alpha}, \tag{23}$$

which is a PH specification, despite the fact that, according to the underlying model, there is unobserved heterogeneity. For example, if the underlying MPH model has a constant baseline hazard $\psi(t) = 1$, then the observed hazard has the (popular) Weibull PH specification with baseline hazard $\alpha t^{\alpha-1}$, with $0 < \alpha = \epsilon < 1$, which displays negative duration dependence[28]. Suppose that $\theta_0(x) = \exp(x'\beta)$. If the true model has a Positive Stable distribution of unobserved heterogeneity and if the researcher assumes instead that there is no unobserved heterogeneity and that $t|x$ has a PH specification (an assumption that is confirmed by the data!) then the parameter of interest β is estimated by $\beta\alpha$, so it is under-estimated in absolute value.

[27] The corresponding densities are bell-shaped [see Hougaard (1986)]. Hougaard (1986) provides a justification of this family as a family of distributions for υ in MPH-type models. Suppose that the individual duration can end for a number of different reasons $\{1, \ldots, n\}$, with cause-specific individual hazards that share the same baseline hazard and the same systematic hazard but not the same individual heterogeneity value υ_j. The individual hazard, which is the sum of the cause-specific individual hazards, then equals $\sum \psi(t)\,\theta_0(x)\,\upsilon_j$, and this is an MPH specification with $\upsilon = \sum \upsilon_j$. Now suppose that the υ_j are i.i.d. positive random variables, and suppose that $n \to \infty$. If the scaled mean of the υ_j has a nondegenerate limiting distribution then it must be a Positive Stable distribution [Feller (1971)]. In fact, for a wide range of distributions of the underlying random variable, the limiting distribution converges to a Positive Stable distribution. So, if υ is an average of many different i.i.d. unobserved heterogeneity terms, then, in many cases, the distribution of υ is approximated by a Positive Stable distribution. Note however that the underlying assumption that the different cause-specific hazards have the same baseline hazard and systematic hazard, while perhaps often reasonable in medical science, is often untenably strong in economics. Moreover, if υ has a Positive Stable distribution and the parameter α is not fixed, then the MPH model is not identified (see below).

[28] If the underlying hazard has Weibull duration dependence $\psi(t) = (1/\alpha)\,t^{1/\alpha-1}$ and G is a Positive Stable distribution with parameter α then the observed hazard does not change with t, so $t|x$ has an exponential distribution.

These results have very important implications. First, *the MPH model is nonparametrically unidentified if the assumption that* $\mathrm{E}(v) < \infty$ *is dropped* (or, alternatively, if Assumption 8b is dropped). Moreover, the adoption of a model that is observationally equivalent to (but different from) the true model leads to biased inference on the parameters of interest [see also Robins and Greenland (1989)]. This is bad news, as it is often difficult to make any justified assumption on the tail of the unobserved heterogeneity distribution. On the other hand, in the case where v represents an important economic variable, economic theory often provides a justification of $\mathrm{E}(v) < \infty$. In Subsection 5.5 we discuss some examples of this.

Ridder (1990) addresses the fundamental identification problem in detail. He argues that for any MPH model with $\mathrm{E}(v) < \infty$ there are observationally equivalent models with $\mathrm{E}(v) = \infty$. In particular, for any MPH model with $\mathrm{E}(v) < \infty$ there is basically one observationally equivalent MPH model satisfying Equation (22), for any $\epsilon \in (0, 1)$. So, Assumption 8 as well as Assumption 8b for given ϵ can all be interpreted as different untestable normalizations that impose identifiability on a class of models that are unidentified.

Let us return to the case where v is degenerate (i.e., the PH model). Van den Berg (1992) proves that the full set of MPH models that is observationally equivalent to the PH model consists of models in which v is degenerate or has a Positive Stable distribution. In the latter case, as is clear from Equation (23), the duration dependence of the baseline hazard and the absolute size of the effect of x are more positive than in the resulting PH model. For the general case, Ridder (1990) shows that some aspects of the MPH model are still identified if no assumptions on the tail of G are made. For example, the sign of the effect of x is identified.

As we shall see below, one solution to the fundamental identification problem is to rely on economic theory when choosing a functional form for G. Another solution is to use information on multiple spells for the same individuals.

5.3. Interaction between duration and explanatory variables in the observed hazard

In this subsection we examine properties of the observed hazard $\theta(t|x)$ if the underlying model has an MPH specification. These provide additional insights into the identification of the model. Throughout most of this subsection we assume that $\mathrm{E}(v) < \infty$, i.e., we adopt the MPH framework of Elbers and Ridder (1982). At times we generalize results by examining the wider class of models where $\mathrm{E}(v) \leqslant \infty$.

If there is no unobserved heterogeneity (so v is a constant), then the observed hazard $\theta(t|x)$ is multiplicative in t and x. Now suppose there is unobserved heterogeneity. If the observed hazard $\theta(t|x)$ would be multiplicative in t and x then the model would be observationally equivalent to a model without unobserved heterogeneity. Because of the nonparametric identifiability of the model, we know that the latter cannot be true. Therefore, the observed hazard cannot be multiplicative in t and x. As a result, we obtain the fundamental insight that *identification of G in MPH* *models comes from nonproportionality of the observed hazard* $\theta(t|x)$ [see Hougaard (1991), Van den Berg (1992) and Keiding (1998)]. In terms of Equation (21): if there is unobserved heterogeneity then the function $h(z(t)\,\theta_0(x))$ is not multiplicative in t and x, and the interaction between t and x identifies G. Yet another way to formulate this is by stating that if there is unobserved heterogeneity then $\log\theta(t|x)$ is not additive in t and x, so for some t and x

$$\frac{\partial^2 \log\theta(t|x)}{\partial t\partial x} \equiv \frac{\partial^2 \log h(z(t)\,\theta_0(x))}{\partial t\partial x} \neq 0, \tag{24}$$

provided that x varies continuously and the appropriate differentiability conditions are satisfied.

Now recall from the previous subsection that if the assumption that $\mathrm{E}(v) < \infty$ is dropped then a proportional specification for $\theta(t|x)$ can also be generated by MPH models with unobserved heterogeneity. Such models are characterized by the property that v has a Positive Stable distribution. All other distributions for v with $\mathrm{E}(v) = \infty$ generate $\theta(t|x)$ that is not multiplicative in t and x. Consequently, if Positive Stable distributions are ruled out for v then the result on the relation between unobserved heterogeneity and nonproportionality of the observed hazard can be extended to include infinite-mean distributions for v.

In fact, unobserved heterogeneity can not generate just any type of interaction between t and x in $\theta(t|x)$. Van den Berg (1992) shows that it is not possible that there are whole intervals of t and x on which there is no interaction[29]. (Whether the interaction is "large" is an empirical matter; as we shall see below, it is not difficult to construct examples in which there is virtually no interaction for a wide range of values of t.) Also, the following simple and appealing specification for $\theta(t|x)$ that allows for interaction cannot be generated with an MPH model,

$$\theta(t|x) = \psi(t)\,\theta_0(x)\,\mathrm{e}^{-\alpha z(t)\,\theta_0(x)},$$

because the function $h(s) = \exp(-\alpha s)$ cannot be generated by the model[30]. In the next subsection we also derive restrictions on the sign of the interaction for different t. All of this evidence implies that the class of models for $\theta(t|x)$ that is generated by MPH models is smaller than the general class of interaction models for $\theta(t|x)$. In other words, the MPH model is overidentified. The fact that the function h must be such that it can be generated by a Laplace transform, the fact that $z(t)$ and $\theta_0(x)$ affect

[29] This follows because any distribution G that gives a function h such that $h(z(t)\,\theta_0(x))$ is multiplicative in t and x on an interval must be a Positive Stable distribution.

[30] This can be seen as follows. If the model is an MPH model then $h(s)$ can be written as $-\mathcal{L}'(s)/\mathcal{L}(s)$, with $\mathcal{L}(s)$ being the Laplace transform of G. However, the function $\mathcal{L}(s)$ that follows from the candidate $h(s) = \exp(-\alpha s)$ is not completely monotone and hence cannot be a Laplace transform [see Feller (1971)].

the value of h only by way of their product, and the fact that t enters the interaction term by way of the integral of the multiplicative term $\psi(t)$, all impose restrictions on $\theta(t|x)$ as a function of t and x.

At this stage it is instructive to examine the results in McCall (1996) on the identification of an extension of the MPH model with $\mathrm{E}(v) < \infty$ and $\theta_0(x) = \exp(x'\beta)$. Specifically, he allows the parameter β to vary with t. This is an empirically relevant extension (recall the discussion at the end of Section 4). Note however that the extension creates a second type of interaction between t and x in the observed hazard, so the question arises whether the data enable a distinction between them. McCall (1996) shows that the model is not identified if x can assume only two different possible values. However, if there is an explanatory variable that attains all possible values between $-\infty$ and ∞ then the model (i.e., ψ, G and $\beta(t)$) is identified, so then the two types of interaction can be distinguished empirically.

The inclusion of time-varying covariates (which is another empirically relevant extension of the MPH model) creates yet another type of interaction between t and x in the observed hazard. It is clear that in some cases a model with time-varying covariates is not identified (for example, if $\theta_0(x(t))$ is multiplicative in t). However, Honoré (1991) illustrates that in some cases time-varying covariates can also be helpful for identification. Suppose that x is time-invariant for part of the population; some of them have the value x_1 while others have x_2, with $\theta_0(x_2) \neq \theta_0(x_1)$. Suppose in addition that for the other part of the population the value of x changes discretely from x_1 to x_2 at duration $t^* > 0$, and assume that x satisfies the conditions for time-varying covariates laid out in Subsection 4.2. Then the model is identified without any assumption on the tail of G (so $\mathrm{E}(v)$ may be finite or infinite). See Heckman and Taber (1994) for a generalization of this result.

The results in McCall (1996), Honoré (1991) and Heckman and Taber (1994) illustrate the fact that the interaction generated by the presence of unobserved heterogeneity is rather specific. It is plausible that as more and more sources of interaction are included into the model, it becomes more and more difficult to achieve identification. In the limit, the assumption that the underlying hazard is multiplicative in t, x and v is essential for identification. If this assumption is dropped then obviously any nonproportional specification can be generated without the need to allow for unobserved heterogeneity, and the model would be unidentified [see also Heckman (1991)]. In particular, the specification (19) can also be generated as an individual hazard, which equals the observed hazard because of the absence of unobserved heterogeneity.

5.4. The sign of the interaction

In this subsection we examine the sign of the interaction between t and x in $\theta(t|x)$. This sign is a potentially interesting model characteristic, as its empirical counterpart may be readily observed from the data. Moreover, economic theory sometimes makes predictions of the sign of the interaction. For example, the ranking model of unemployment by Blanchard and Diamond (1994) predicts that the aggregate exit rate out of unemployment as a function of t decreases more in a "bad" steady state (i.e., a steady state where the exit rates are low anyway) than in a good steady state. If the steady state is represented by a dummy variable x then this means that the interaction between t and x is predicted to be always positive.

The discussion is facilitated by using $\theta_0(x)$ and x interchangeably. Obviously, this entails no loss of generality in the examination of the sign of the interaction, provided that it is kept in mind that x has a *positive* effect on $\theta(t|x,\upsilon)$. For convenience we take x to vary continuously, so that the sign of the interaction can be expressed as the sign of the cross-derivative of $\log h(z(t)x)$ with respect to t and x [see Equation (24); recall that $\theta(t|x) = \psi(t) \cdot \theta_0(x) \cdot h(z(t)\,\theta_0(x))$].

The derivative of $\log h(z(t)x)$ with respect to x equals $h'(z(t)x)z(t)/h(z(t)x)$. The sign of the cross-derivative of $\log h(z(t)x)$ with respect to t and x then equals the sign of the derivative of $sh'(s)/h(s)$ evaluated at $s = z(t)x$. The function $h(s)$ is determined by the Laplace transform $\mathcal{L}(s)$ of G. Therefore, the sign of the interaction at a certain t and x is completely determined by G[31]. Given that $z(t)x$ takes on all values in $[0,\infty)$, knowledge of the sign of $sh'(s)/h(s)$ for all s is necessary in order to infer whether this sign is unambiguous for all t and x. To put this more bluntly, the full specification of the unobserved heterogeneity distribution determines the sign of the interaction between duration and explanatory variables in the observed hazard.

The first notable result concerns the sign of the interaction for small t. In general, the interaction is strictly negative on an interval $[0,\varepsilon)$[32]. This negative interaction means that if x is large then the observed duration dependence for small t is more negative than if x is small. This can be understood as follows. In the sub-population of individuals with a high value of x, the individuals who also have a high υ will have a disproportionally high hazard. As a result, those individuals leave the state very quickly, and this has a strong negative duration-dependence effect on the observed hazard for the individuals with high x. Among the individuals with low x, this weeding out phenomenon occurs at a much lower speed, so their observed hazard decreases less strongly. It is important to stress that this intuitive explanation does not work for $t > 0$, because the distribution of υ among survivors at $t > 0$ depends on x itself.

Lancaster (1979) shows that if G has a Gamma distribution,

Family of Gamma distributions:

$$g(\upsilon) = c^r/\Gamma(r) \cdot \upsilon^{r-1} \exp(-c\upsilon) \quad \text{for all } \upsilon > 0, \text{ with } c, r > 0,$$

then the interaction is negative for all t and x, so the negative interaction sign for small t can be extended to all t. Unfortunately, this result cannot be generalized to include

[31] It follows from the results in Subsection 5.1 that $sh'(s)/h(s)$ at $s = z(t)x$ can be expressed in terms of the moments of $\upsilon|T > t, x$ (specifically, it depends on the first three moments).

[32] For example, if $\mathrm{Var}(\upsilon) < \infty$ and $\lim_{t\downarrow 0} \psi(t) \in (0,\infty]$ then $\partial^2 \log \theta(t|x)/\partial t \partial x < 0$ at $t = 0$. If $\mathrm{E}(\upsilon^3) < \infty$ and $\lim_{t\downarrow 0} \psi(t) \in [0,\infty]$ then $\partial^2 \log \theta(t|x)/\partial t \partial x < 0$ on an interval next to $t = 0$.

all possible G. To see this, consider discrete distributions for G with a finite number of mass points (or points of support), each of them positive and finite,

Family of discrete distributions with a finite number of mass points,
each of them positive and finite:

$$\Pr(v = v_i) = p_i \quad \text{for all } i = 1, 2, \ldots, n, \text{ with}$$
$$0 < v_1 < v_2 < \cdots < v_n < \infty, \quad 0 < p_1, p_2, \ldots, p_n < 1, \quad \sum\nolimits_{i=1}^{n} p_i = 1, \ n < \infty$$

(this is a popular specification in empirical work; see Subsection 5.5 below). We shall show that it is intuitively plausible that in this case, as $t \rightarrow \infty$, the derivative $\partial \log \theta(t|x)/\partial x$ goes to its value at $t = 0$ (so that this derivative varies with t in a non-monotone way, i.e., the cross-derivative does not have the same sign everywhere). When t increases, the group of survivors becomes increasingly more homogeneous, since the individuals with $v > v_1$ leave unemployment on average earlier than the individuals with $v = v_1$. In the limit, the group of survivors is homogeneous (all remaining individuals have $v = v_1$) so the value of $\partial \log \theta(t|x)/\partial x$ equals the value in a model without unobserved heterogeneity, which is $\theta_0'(x)/\theta_0(x)$ (see Equation 20). This in turn equals the value that is taken by $\partial \log \theta(t|x)/\partial x$ in general at $t = 0$ (see Equation 20), because at $t = 0$ the selection due to heterogeneity has not yet taken place[33].

Example 7. Let v have a discrete distribution with two points of support with $\Pr(v = \frac{1}{5}) = \Pr(v = \frac{3}{5}) = \frac{1}{2}$. Then the cross-derivative of $\log \theta(t|x)$ with respect to t and x equals zero if $z(t)\,\theta_0(x)$ is about 4.6 and it is positive if and only if $z(t)\,\theta_0(x)$ exceeds that number.

In this example, there is a positive value of $z(t)\,\theta_0(x)$ for which the observed hazard is multiplicative in t and x (i.e., the cross-derivative is zero) despite the presence of unobserved heterogeneity. However, the corresponding values of t and x have measure zero in the set of all possible values of t and x. Note that the above result implies that, if G is discrete with a finite number of points of support, the observed hazard $\theta(t|x)$ can be approximated by a PH specification if t is sufficiently large.

Incidentally, it is not difficult to construct examples where the weeding out of individuals with high v occurs very quickly after $t = 0$. If v has two points of support where one of them is extremely large, then the individuals with large v leave the state almost immediately. As a result, the *magnitude* of the interaction between x and t is virtually zero for almost all $t > 0$.

[33] These results imply that, when comparing an individual with a relatively small x to one with a relatively large x, the proportionate difference between the observed hazards diminishes as time starts to run from $t = 0$ onward, but it ultimately returns to the level at $t = 0$.

The family of discrete distributions is not the only family that generates a non-monotone sign of the interaction. Other examples include uniform distributions with support $[c_1, c_2]$ with $0 < c_1 < c_2 < \infty$ as well as many other distributions with a positive lower bound of the support [see Abbring and Van den Berg (2001) for details]. In general, it seems difficult to derive conditions on G such that the interaction is always negative[34]. In the next subsection we return to this issue, when we examine the limiting distribution of $\upsilon | T > t, x$ as $t \to \infty$, for a wide class of distributions G.

Recall that in general for small t the interaction is negative. It turns out that, even if the interaction may be positive for larger t, the *cumulative* interaction remains negative. With this we mean that (under suitable regularity conditions),

$$\int_0^t \frac{\partial^2 \log \theta(\tau | x)}{\partial \tau \partial x} \, \mathrm{d}\tau < 0,$$

for all t and x. This can be seen by noting that this integral equals $\partial \log \theta(\tau|x)/\partial x$ at $\tau = t$ minus the same expression at $\tau = 0$, and, by Equation (20), this is negative.

We end this subsection by noting a remarkable result on the effect of x on the observed hazard $\theta(t|x)$ in MPH models[35]. One may be tempted to think that this effect is always positive if x has a positive effect on the underlying hazard $\theta(t|x, \upsilon)$. However, this is not a general property of the model. Intuitively, if a fraction of individuals has a very high value of υ then, in the sub-population of individuals with high x, the high-υ individuals leave the state extremely quickly. The drop in the mean value of υ among the survivors with high x is then so large that their hazard may on average fall below the value of those with lower x values. In such a case, the negative effect of the drop in υ on $\theta(t|x)$ is not offset by the positive effect of the large x. In terms of Equation (20), the second term on the right-hand side dominates the first one.

Example 8. Consider again the discrete distribution for υ with $\Pr(\upsilon = \frac{1}{5}) = \Pr(\upsilon = \frac{3}{5}) = \frac{1}{2}$ (see Example 7). Then $\partial \theta(t|x)/\partial x$ is always positive. However, if the highest mass point is at $\frac{5}{2}$ instead of $\frac{3}{5}$ this derivative is negative for values of t and x such that $z(t)\theta_0(x)$ is in an interval around 1.

In sum, the observed hazard of a high-x individual can be smaller than that of a low-x individual. This means that it is not possible to deduce the sign of the effect of x on the underlying individual hazard from the observed relation between x and the observed hazard at a certain duration t. It should however be stressed that this remarkable effect can only occur for some local duration intervals. Specifically, the observed survivor

[34] Negative interaction is equivalent to the statement that $-\mathcal{L}'(e^y)/\mathcal{L}(e^y)$ is log-concave on $y \in (-\infty, \infty)$, but this does not seem to correspond to a well-known class of distributions for G.

[35] Even though this result is not concerned with the sign of the interaction, its interpretation fits in with the latter subject.

function $\overline{F}(t|x)$ and the observed mean duration $\mathrm{E}(t|x)$ are always decreasing in x (iff $\theta_0(x)$ increases in x). This can be seen from the relations

$$\overline{F}(t|x) = \mathrm{E}_\upsilon(\overline{F}(t|x,\upsilon)) = \mathcal{L}(z(t)\,\theta_0(x)),\ \ \mathrm{E}(t|x) = \mathrm{E}_\upsilon \mathrm{E}(t|x,\upsilon) = \int_0^\infty \mathcal{L}(z(t)\,\theta_0(x))\,\mathrm{d}t,$$

where E_υ denotes the expectation with respect to G (note that $\mathcal{L}$ decreases in its argument; see Equation 18).

5.5. *Specification of the unobserved heterogeneity distribution*

Studies in which parameterized MPH models are estimated have wrestled with the choice of a functional form for G [see e.g., Heckman and Singer (1984a)]. This choice is thought to be harder to justify than the choice for a functional form for the baseline hazard ψ, as economic theory often suggests a shape for the latter. In this subsection we examine parametric families of distributions that can be given supporting arguments as a choice for G. We start with families that can be supported by limit arguments. Next we show that economic theory sometimes actually does make informative predictions on important aspects of the shape of G. This typically concerns cases where a key source of individual heterogeneity is observed by labor market participants but not by the researcher.

5.5.1. *Discrete distributions*

Suppose that the baseline hazard and the systematic hazard have parametric functional forms with a finite number of parameters, but that the only assumption on G is that it has a finite mean (or satisfies Equation 22). For this case, Heckman and Singer (1984c) show that the Maximum Likelihood estimator of G is a discrete distribution, provided that some regularity conditions are met[36]. For a given sample, the parameters of this discrete distribution (the number of points of support, their location, and their associated probabilities) are chosen such as to maximize the likelihood function. The result by Heckman and Singer (1984c) illustrates the flexibility of discrete distributions as heterogeneity distributions. Intuitively, if the number of points of support increases, then any true underlying distribution G can be approximated well. In practice, it is often difficult to find more than a few different mass points. Usually, if more than two or three points of support are taken then the estimates of some of them coincide. Standard practice in case of discrete G is to estimate the model with a number of mass points that is either predetermined or equal to the maximum number that could be detected, and to report standard errors conditional on this choice. It is important to stress that such

[36] See Trussell and Richards (1985), Lancaster (1990) and Baker and Melino (2000) for additional insights into this estimator and for alternative computational strategies.

approaches are not “nonparametric” in the true sense of the word, and that the standard errors do not reflect uncertainty with respect to the actual number of mass points.

The fact that it is often difficult to find more than a few mass points may reflect a lack of informativeness on G in the data. Recall that the data do not provide observations on drawings from G, but that G enters the likelihood function as a mixing distribution. The information on G comes from the observed interaction between t and x in the data, and it may be that a mixing distribution with a few mass points is often able to capture most of this. The simulations in Heckman and Singer (1984c) strongly confirm this. They find that the parameters of ψ and θ_0 as well as the shape of the distribution of $t|x$ are well estimated if G is assumed to be discrete with an unknown number of mass points, even if the true G is continuous. The estimated number of mass points is typically small.

For G discrete with a finite number of points of support, each of them positive and finite, we restate the following model properties. First, $\mathrm{E}(\upsilon) < \infty$. Secondly, the interaction between t and x in $\theta(t|x)$ is not monotone; it is negative for small t and positive for very large t. Thirdly, the effect of x on $\theta(t|x)$ is not always monotone even if the effect on $\theta(t|x, \upsilon)$ is.

5.5.2. Gamma distributions

In applications, the family of Gamma distributions has perhaps been the most popular choice for G. This stems from the resulting analytic tractability: all relevant properties of the distribution of $t|x$ can be expressed in closed-form solutions. In their recent working paper, Abbring and Van den Berg (2001) are the first to provide a less ad-hoc justification for the choice of the family of Gamma distributions for G. Suppose that zero is the lower bound of the support of the true (unknown) G, with υ being a continuous random variable (we do not make assumptions on the upper bound of the support of G). Then, under mild regularity conditions, the unobserved heterogeneity distribution among the survivors at duration t converges to a Gamma distribution if $t \to \infty$. In fact, we have to scale the distribution of υ among survivors because the unscaled distribution converges to zero (note that the Gamma family is invariant to scaling). This result implies that, in many cases, the heterogeneity distribution among survivors at high durations can be approximated well by a Gamma distribution, and this provides a motivation to adopt the Gamma family for $G(\upsilon)$ itself.

For $G(\upsilon)$ equal to a Gamma distribution, we restate the following model properties. First, $\mathrm{E}(\upsilon) < \infty$. Secondly, the interaction between t and x in $\theta(t|x)$ is monotone and negative for all t. Thirdly, the effect of x on $\theta(t|x)$ is always monotone if the effect on $\theta(t|x, \upsilon)$ is monotone.

The limit result in Abbring and Van den Berg (2001) does not hold if the true $G(\upsilon)$ is a discrete distribution with a finite number of points of support[37].

[37] Recall that in such a case the sign of the interaction is positive for large t, whereas in the case of a Gamma distribution it is negative for large t. The latter suggests that, if in practice a choice must be

5.5.3. Suggestions from economic theory

Now let us turn to (aspects of) shapes of $G(v)$ that can be justified by economic theory. First, as a general remark, it should be noted that economic theory often predicts that the exit rate out of a state is bounded from above. Consider the search theories of Section 3. In general, the exit rate out of unemployment can be written as $\lambda\overline{F}(\phi)$. The second term in this expression is a probability which necessarily lies between zero and one. If the first term is infinite then there are no frictions in the first place, and the models reduce to standard labor market models with zero unemployment durations. According to this line of reasoning, $\theta(t|x,v)$ should be bounded from above, which implies that the support of G is bounded from above (which in turn implies that $\mathrm{E}(v) < \infty$)[38].

5.5.3.1. Suggestions from equilibrium search models. Suppose worker behavior is described by the search models of Section 3. In the literature, these models have been extended to include employer behavior. For surveys of the theoretical and empirical analysis of such "equilibrium search models", see Ridder and Van den Berg (1997), Mortensen and Pissarides (1999) and Van den Berg (1999). To fix thoughts, consider the equilibrium search model of Bontemps, Robin and Van den Berg (1999) where unemployed and employed workers search, and different workers have different values of leisure b. If the job offer arrival rates are the same in employment and unemployment, then the reservation wage of an unemployed worker with value of leisure b is simply equal to b. Now suppose that b has a continuous distribution $H(b)$ in the population. An employer sets his wage w such as to maximize his steady-state profits. We assume that the number of firms is fixed, or, alternatively, that an entry fee has to be paid. It is not optimal for any firm to offer a wage equal to the lower bound $\underline{b}$ of the distribution $H(b)$, because then its steady-state labor force and profit rate are zero. The lowest wage $\underline{w}$ in the market is strictly larger than $\underline{b}$. As a result, there is a positive fraction of individuals who accept any wage offer (i.e., who have $b < \underline{w}$).

In this model, the individual exit rate out of unemployment equals $\lambda\overline{F}(b)$. Now suppose that the researcher wants to estimate a reduced-form model of unemployment durations. The individual value of leisure b is unobserved, so it is reasonable to take the unobserved heterogeneity term v to represent the acceptance probability $\overline{F}(b)$ (provided that there is no additional source of unobserved heterogeneity). As a result,

made between a discrete G or a Gamma G, it is useful to examine the sign of the interaction between t and x in the data on $\theta(t|x)$ for large t [see Hougaard (1991) for an example].

[38] One may argue that λ is affected by an optimally chosen search intensity, and that the distribution of structural determinants in the population is such that the resulting distribution of λ does not have an upper bound. However, in search and matching models, λ is at least partially determined by the meeting technology of the labor market; this technology is a market characteristic that cannot be fully dominated by individual behavior.

the distribution $G(\upsilon)$ has support in $[0, 1]$. But there is a positive fraction of workers with $\overline{F}(b) = 1$, so G has a mass point at the upper bound of its support (i.e., at $\upsilon = 1$). If the highest wage in the market $\overline{w}$ is smaller than the highest level of b then G also has a mass point at zero. In that case G is a defective distribution; a positive fraction of individuals is unemployed forever. In practice it may not be difficult to sort out the latter individuals from the data (i.e., to observe whether $b > \overline{w}$), because it does not make sense for these individuals to search for a job, so they may classify themselves as being nonparticipants.

It is not difficult to see that this result extends to more general equilibrium search models. Often, employer behavior is such that a positive fraction of unemployed workers accepts any wage offer and consequently has the maximum hazard level for the transition into employment.

5.5.3.2. Suggestions from on-the-job search models. Consider the stationary on-the-job search model of Subsection 3.2. Published statistics on nationwide job mobility contain information on the marginal job duration distribution, i.e., on the distribution of job durations unconditional on the wage in the job. The wage then represents unobserved heterogeneity in the job duration data.

The distribution of t given the wage w on the job is exponential with density

$$f(t|w) = (\delta + \lambda_1 \overline{F}(w))\, \mathrm{e}^{-(\delta + \lambda_1 \overline{F}(w))\, t}. \tag{25}$$

Consider the job durations t of a cohort of workers who have just left unemployment for a job (this constitutes the inflow into employment at a given point of time). If all unemployed workers accept any wage that is offered to them then, in this cohort, the wage w is distributed according to $F(w)$. To obtain the marginal job duration distribution for this cohort, we have to integrate Equation (25) with respect to $\mathrm{d}F(w)$. This gives

$$f(t) = \frac{1}{\lambda_1} \int_{\delta}^{\delta + \lambda_1} z\, \mathrm{e}^{-zt}\, \mathrm{d}z,$$

which is a "mixture of exponentials", i.e., a mixture of distributions with constant hazards, with a uniform mixture distribution for the hazards with support on the interval $(\delta, \delta + \lambda_1)$[39]. This is not surprising. The conditional hazard of $t|w$ is constant over the job duration. It is then mixed with respect to a determinant (w) of the

[39] This can be further simplified to

$$f(t) = \frac{\mathrm{e}^{-\delta t}}{\lambda_1 t^2} \left[1 + \delta t - (1 + (\delta + \lambda_1)\, t)\, \mathrm{e}^{-\lambda_1 t} \right].$$

conditional hazard. Workers are merely concerned with the ordering of the current wage and the wage offer, and not with the shape of the underlying wage offer distribution itself. Their location on the job ladder therefore determines their hazard. Note that, as a result, the marginal job duration distribution does not depend on F.

In terms of an MPH model, $\theta(t|x)$ can be thought of as being generated by $\theta(t|x, v) = v$, where v has a uniform distribution on $(\delta, \delta + \lambda_1)$[40]. This result for a cohort of newly employed workers can be generalized to other (more relevant) sampling schemes. Ridder and Van den Berg (1998) apply this approach to study job mobility with aggregate data.

The argument above also applies to other settings where only the rank of the individual's heterogeneity value affects the individual's hazard rate, and where these values and their ranks are unobservable. Moscarini (1997) examines a job search model for the unemployed where individuals are ranked by employers on the value of some time-invariant characteristic. The rate at which an individual obtains a job depends on the fraction of the unemployed that has worse characteristics. For a specific matching technology, this results in an unemployment duration distribution that is again a mixture of exponential distributions with a uniform mixture distribution.

5.6. Effects of misspecification of functional forms

Generally, in applications, ψ and/or $G(v)$ are assumed to have a parametric functional form [see Lancaster (1990) for a catalogue of popular functional forms]. We finish this section on properties of the MPH model by summarizing some results on the effects of misspecification of these functional forms on the probability limits of the Maximum Likelihood (ML) estimates. Throughout the subsection (and in line with this literature) we assume that

$$\theta_0(x) = \exp(\beta_0 + x'\beta_1),$$

and that all moments of v exist. The model is normalized by taking $\mathrm{E}(v) = 1$. The only type of censoring that is considered concerns independent right-censoring at a fixed duration.

A natural starting point concerns the misspecification due to omission of unobserved heterogeneity from the model, if it is present in the data-generating process. Recall that in Subsection 5.1 we argued that the estimated duration dependence will be too negative, and the effect of x may be inconsistently estimated as well. Gail, Wieand and Piantadosi (1984) provide the following result. If the baseline hazard $\psi(t)$ is known a priori, if one erroneously ignores unobserved heterogeneity in the model

[40] Note that if δ or λ_1 depend on t or x, then this is not an MPH model anymore.

specification, *and* if there is no censoring, then β_1 is consistently estimated with ML. In fact, it is not difficult to show that

$$\text{plim}\,\widehat{\beta_0} = \beta_0 - \mathrm{E}\left(\frac{1}{v}\right) < \beta_0, \qquad \text{plim}\,\widehat{\beta_1} = \beta_1,$$

where $\text{plim}\,\widehat{\beta_i}$ denotes the probability limit of the ML estimator of β_i (i.e., the value to which the estimate converges in probability as the sample size increases). Note that $\mathrm{E}(1/v) > 1/\mathrm{E}(v) = 1$ if and only if $\mathrm{Var}(v) > 0$, i.e., if there is unobserved heterogeneity[41, 42].

Unfortunately, these welcome results do not generalize in any way to more realistic settings. Ridder (1987) shows that censoring in the data makes $\widehat{\beta_1}$ inconsistent (unless the specified G equals the true G or $\beta_1 = 0$). The asymptotic bias is towards zero if the specified model assumes absence of unobserved heterogeneity. Lancaster (1985b) shows that if the baseline hazard is known to have a Weibull specification with an unknown parameter, one ignores unobserved heterogeneity, and there is no censoring, then the estimates of both the Weibull parameter and β_1 are asymptotically biased towards zero. In fact, they are all biased in the same proportion. Basically, in this case, ML gets the regression function for $\log t$ right, but we are after the original parameters of the individual hazard function instead of the elasticities of the mean log duration. Ridder (1987) also shows that misspecification of the shape of the baseline hazard results in inconsistency of $\widehat{\beta_1}$.

The results above are all analytically derived. For more general model settings, the effects of misspecification have been analyzed by way of extensive Monte Carlo simulations. Ridder (1987) allows for censoring in the Lancaster (1985b) model, and he allows for misspecified G in the assumed model. It turns out that censoring exacerbates the asymptotic bias in $\widehat{\beta_1}$ due to misspecification of G, and the results become sensitive to the assumed specification of G. Moreover, it turns out that the estimates display a large small-sample bias even if the model specification is correct. This bias disappears very slowly when the sample size increases. Such small-sample biases are absent for the PH model without unobserved heterogeneity; see Andersen, Bentzon and Klein (1996).

[41] See also Lancaster (1983). Ridder (1987) generalizes this result by proving the following: if the baseline hazard is known in advance, the assumed G is fully specified without unknown parameters, the assumed G is not equal to the true G, and there is no censoring, then β_1 is consistently estimated.

[42] This is not in conflict with the result in Subsection 5.1 that $d\log\theta(t|x)/dx = \beta_1(1-a)$ for some $a > 0$. Somewhat loosely one may say that $\widehat{\beta_0}$ ensures that the average level of the specified $\log\theta(t|x)$ agrees to the average level in the data, and that the effect of x in the data is best captured by $\widehat{\beta_1} = \beta_1$. Note that in this specific model, $\mathrm{E}(\log z(t)|x, v)$ is additive in v and x. In particular, $\mathrm{E}(\log z(t)|x, v) = -\beta_0 - x'\beta_1 - \log v + c$, with $c \approx -0.58$ being the mean of an EV1 random variable, and with the function $z(.)$ completely known. So by analogy to the regression model, dispersion in v does not affect the estimate of β_1.

Ridder (1987) also examines the performance of ML estimation of an assumed model with a Weibull baseline hazard and a Gamma distribution for v, if both are misspecified. The simulations reinforce the negative results above. Ridder (1987) conjectures that if the baseline hazard is flexibly specified with a sufficient number of unknown parameters, and if censoring is virtually absent, then it does not matter which family of distributions is assumed for G in order to obtain a reliable estimate of β_1. However, the simulation results in Baker and Melino (2000) go against this[43]. Most of the biases due to the above problems can be substantial, depending on the situation at hand. For the Partial Likelihood estimation method, similar results have been derived [see e.g., Bretagnolle and Huber-Carol (1988)].

By now there are also many studies of real-life single-spell data in which it is reported that the estimates of (the parameters of) β_1, ψ and G are sensitive to changes in the assumed family of distributions for G or the assumed set of x or the assumed functional form of ψ, even though sometimes the over-all fit of the model does not change with this in any substantial way [see e.g., Heckman and Singer (1984a), Trussell and Richards (1985), Hougaard, Myglegaard and Borch-Johnsen (1994)]. Keiding, Andersen and Klein (1997) provide a survey of studies with biostatistical data.

The recent literature on semiparametric and nonparametric estimation of the MPH model provides some interesting additional insights on this. First of all, Hahn (1994) examines models with Weibull duration dependence, and he assumes that v is a continuous random variable with a finite mean. He shows that with single-spell data, the information matrix is singular, and that there is no $\sqrt{n}$-consistent estimator for β_i and the Weibull parameter[44]. Thus, in a certain sense, there is less information on the model parameters than what is typically available in econometric analyses. Secondly, Heckman and Taber (1994) and Kortram et al. (1995) show that the mapping from the data-generating process to the data is not continuous, so that two distinct MPH models can generate very similar data[45]. Thirdly, the nonparametric (or semiparametric) estimator developed by Horowitz (1999) has convergence rates that are smaller than $\sqrt{n}$. In particular, under certain assumptions (including absolute continuity of an element of x, differentiability of $\psi(t)$ and the density of v, and $\mathrm{E}(v^2) < \infty$), the convergence rates of β_i and ψ can be at most almost equal to $n^{-2/5}$, which is obviously slower than $n^{-1/2}$. For the heterogeneity distribution and density G and g, the rate of convergence is $(\log n)^{-2}$, which is *very* slow.

[43] It should be noted, though, that Baker and Melino (2000) do not examine an MPH model but a discrete-time model where the individual per-period exit probability is a logistic function of $\psi(t)\theta_0(x)v$. Whether these models behave similarly is an issue for further research.

[44] See Klaassen and Lenstra (1998) for a generalization of this result.

[45] As an example, consider the simplest MPH model, with $\theta_0(x) = \exp(x)$ where x is a single dummy variable, and with absence of duration dependence and unobserved heterogeneity. The distribution of $t|x$ is virtually the same as the distribution generated by an MPH model with $\theta_0(x) = \exp(2x)$, duration dependence proportional to $2t$, and v distributed as a Positive Stable distribution with parameter $\frac{1}{2}$ with the upper tail replaced by a finite mass point [see Kortram et al. (1995) for details; note the similarity to the example in the discussion in Subsection 5.2; also note that here $\mathrm{E}(v) < \infty$].

Together, these results lead to the following conclusion. In the absence of strong prior information on the determinants of the MPH model, single-spell data do not enable a robust assessment of the relative importance of these determinants as explanations of random variation in the observed durations (even if the unobserved heterogeneity mean is known to be finite). Minor changes in the assumed parametric specification, leading to a similar over-all fit, may produce very different parameter estimates. This implies that estimation results from single-spell data are sensitive to misspecification of the functional forms associated with these determinants. Therefore, interpretations based on such results are often unstable and should be performed with extreme caution.

In biostatistics, this state of affairs has led to a renewed interest in Accelerated Failure Time models for the analysis of single-spell duration data [see Hougaard, Myglegaard and Borch-Johnsen (1994) and Keiding, Andersen and Klein (1997) for a survey]. Note that such models allow for robust inference on the effect of x on the mean of $\log t$[46]. In a way, the choice for the AFT model means that all hope is given up on the attempt to (i) disentangle genuine duration dependence from the effect of unobserved heterogeneity, and (ii) quantify the effect of covariates on the *individual* hazard as opposed to the observed hazard, with single-spell data. From an economic-theoretic point of view, however, the AFT approach is unsatisfactory, because, as we have seen in Sections 2 and 4, the parameters of the individual hazard are the parameters of interest. It may therefore be better to exploit predictions from the underlying economic theory when specifying the duration model, and/or look for data with multiple spells[47].

If one is only interested in the *sign* or *significance* of a covariate effect on the individual durations then the AFT approach may be useful. Recall from Subsection 5.4 that in MPH models the sign of the effect of x on the mean duration is always the same as the sign of the effect on the individual hazard, regardless of the specification of ψ or G. Regression of $\log t$ on x therefore provides robust evidence on this sign [see Solomon (1984) for proofs; Li, Klein and Moeschberger (1993) provide supporting Monte Carlo evidence on the performance of test statistics for the significance of the effect of x]. Such an approach may be useful if one is interested in whether participation in a treatment program (to be represented by x) has any effect. However, in economics, data on treatment effects are usually non-experimental and treatment assignment is selective, so then x is not exogenous (see Subsection 9.2).

[46] Indeed, Horowitz (1996) shows that the β parameters in the transformation model (9) can be consistently estimated with an estimator with convergence rate equal to $n^{-1/2}$. Recall that the AFT model is a special case of the transformation model.

[47] Another approach would be to estimate the model nonparametrically using methods described in Subsection 5.2. It is still too early to assess whether this approach is fruitful. Yet another approach is to use population data (if available). See Van den Berg and Van Ours (1996) for an example of this based on a discrete-time model.

6. The MPH model with multi-spell data

6.1. Multi-spell data

This section deals with identification of the MPH model if the data provide durations of multiple spells in a given state by a given individual, i.e., if the data are *multi-spell* data. Here, an individual has a given value of v, and his spell durations are independent drawings from the univariate duration distribution $F(t|x, v)$, where, of course, v is unobserved, so that the durations given just x are not independent. We mostly focus on an "ideal" case in which the data consist of a random sample of individuals and provide two uncensored durations for each individual in the sample. Actually, the use of the term "individual" is not very appropriate here, as the setup includes cases in which physically different individuals are assumed to share the same value of v and we observe one or more durations for each of these individuals. It is convenient to refer to such a group of individuals as a *stratum*. It depends on the context whether one may assume that v, ψ and θ_0 are identical across durations for the same individual or stratum. In subsequent sections we examine more general models, in which ψ and θ_0 may vary across spells, the values of v in different spells may be stochastically related, and other dependencies between the durations are allowed. It is useful to think of the present section as being concerned with a model for a single type of duration, where we have multiple spells of this type of duration for each "individual", whereas the subsequent sections are concerned with models for different types of durations with single or multiple spells of each type for each "individual".

The empirical analysis of MPH models with multi-spell duration data is widespread. For example, Newman and McCullogh (1984) use such data to estimate reduced-form models for birth intervals, while Ham and Rea (1987) and Coleman (1990) use such data to estimate reduced-form unemployment duration models[48]. Lillard (1993) and Lillard and Panis (1996) estimate marriage duration models with multi-spell data. In these applications, the multiple spells with a given value of v are associated with a single physical individual. There are also many applications in which multiple spells with a given v are associated with different physical individuals [see e.g., Kalbfleisch and Prentice (1980)]. The heterogeneity term is then assumed to be identical across individuals within some group or stratum. Typically, different individuals within a stratum are allowed to have different values of x. As we shall see below, this may actually be very useful for inference[49]. Recent applications include Guo and Rodríguez (1992), Wang, Klein and Moeschberger (1995), Sastry (1997), Ridder and Tunalı (1999) and Lindeboom and Kerkhofs (2000). Arroyo and Zhang (1997) survey applications in the analysis of fertility. In studies on lifetime durations of identical twins, the unobserved heterogeneity terms are often assumed to capture

[48] Ham and Rea (1987) use a discrete-time model.

[49] Indeed, with stratified partial likelihood inference, estimation of the systematic hazard θ_0 is *driven* by the variation in x (see Subsection 6.2).

unobserved genetic determinants, so then υ is identical within twin pairs [see e.g., Hougaard, Harvald and Holm (1992a)].

To proceed, note that the individual hazard function $\theta(t|x,\upsilon)$ is the same for both durations associated with the "individual". The value of x may differ between the corresponding spells. If necessary we denote the values by x_1 and x_2, respectively. Conditional on x and υ, the two durations t_1 and t_2 are independent. Conditional on x, the variables t_1 and t_2 are independent if there is no unobserved heterogeneity, i.e., if υ is not dispersed.

If $\theta_0(x) = \exp(x'\beta)$ then

$$\begin{aligned} \log \int_0^{t_1} \psi(u)\,\mathrm{d}u &= -x_1'\beta - \log \upsilon + \varepsilon_1, \\ \log \int_0^{t_2} \psi(u)\,\mathrm{d}u &= -x_2'\beta - \log \upsilon + \varepsilon_2, \end{aligned} \tag{26}$$

where ε_1 and ε_2 are i.i.d. EV1 distributed. Equations (26) suggest a similarity to standard panel data models with fixed effects. We return to this below.

The joint density $f(t_1,t_2|x)$ of t_1 and t_2 given x can be expressed as

$$f(t_1,t_2|x) = \int_0^{\infty}\int_0^{\infty} f(t_1|x_1,\upsilon)f(t_2|x_2,\upsilon)\,\mathrm{d}G(\upsilon), \tag{27}$$

in which G denotes the joint distribution of υ across "individuals" in the population. The density $f(t_i|x_i,\upsilon)$ can of course be expressed in terms of the determinants of θ (see Section 2). The joint survivor function of t_1 and t_2 given x can then be expressed as

$$\overline{F}(t_1,t_2|x) = \int_0^{\infty} \mathrm{e}^{-[z(t_1)\,\theta_0(x_1) + z(t_2)\,\theta_0(x_2)]\upsilon}\,\mathrm{d}G(\upsilon).$$

In many applications, the individual likelihood contribution is based on the density (27). In terms of panel data analysis, this means that the values of υ are treated as "random effects" when estimating the model with Maximum Likelihood[50]. An alternative empirical approach treats υ as individual-specific parameters or "incidental" parameters. The likelihood function is then written for given unknown values of these (and the other) parameters[51].

6.2. Identification results

One may distinguish between two approaches in the literature on identification of the MPH model with multi-spell data. The first approach below is concerned with the full

[50] Here, as in the model with single spells, standard maximization of the likelihood may be computationally unfeasible for particular parametric specifications for G and ψ. In such cases, use of the EM algorithm may be preferable [see Lancaster (1990) for details].

[51] See Lancaster (2000a) for a general overview of incidental parameters in econometrics.

identification of the model and relies on results that were discussed in Section 5. The second approach is concerned with the identification of the systematic hazard θ_0 and follows from properties of a particular estimation method.

We start with the first approach. Honoré (1993) shows that the MPH model with multi-spell data is identified under much weaker assumptions than in Section 5. In fact, we do not need to assume that there are observed explanatory variables x at all. In other words, the analysis is conditional on a given value of x, and we may allow for full interaction of the actual value of x with the model determinants: ψ may depend on x in an unspecified way, and v and x may be dependent in the population. Note that here x does not vary across spells for a given individual. We may write

$$\theta(t|x,v) = \psi(t|x)\cdot v, \qquad v|x \sim G(v|x).$$

This includes of course as a special case that $\psi(t|x)$ can be written as $\psi(t)\,\theta_0(x)$.

This model is identified given regularity assumptions corresponding to Assumptions 2–4, and given a normalization of the integrated baseline hazard (analogical to Assumption 7). Thus, if two observations are available for each v, then the identification of the model does not require an untestable assumption on the tail of the unobserved heterogeneity distribution G anymore, and, perhaps even more importantly, v and x are allowed to be dependent. The identification of this distribution does not come anymore from the interaction between the duration and the observable explanatory variables in the observed hazard. The identification does however need proportionality of the duration effect and the unobserved heterogeneity term in the individual hazard. It should be noted that this model is nevertheless overidentified; see Subsection 8.2.2.

Example 9. Let $\psi = 1$ (so there is no duration dependence) and $x_1 = x_2\,(= x)$, and suppose that v has a Positive Stable distribution (see Subsection 5.2). Such distributions have infinite means. As we have seen, the resulting MPH model for single spells is observationally equivalent to a PH model without unobserved heterogeneity and a Weibull baseline hazard. However, it is easy to see that the joint survivor function of t_1 and t_2 equals

$$\overline{F}(t_1,t_2|x) = \exp\left(-[\theta_0(x)]^{\alpha}(t_1+t_2)^{\alpha}\right)$$

(with $0 < \alpha < 1$), whereas if there is no unobserved heterogeneity and the baseline hazard has a Weibull specification ($\psi(t) = \alpha t^{\alpha-1}$) then

$$\overline{F}(t_1,t_2|x) = \exp\left(-[\theta_0(x)]^{\alpha}(t_1^{\alpha}+t_2^{\alpha})\right),$$

so the two models are observationally distinct, even if $\theta_0 = 1$.

Now let us turn to the second approach to identification, which focuses on the effect of observed explanatory variables on the individual hazard function. The systematic hazard θ_0 is identified under very weak conditions if the data contain multiple spells with the same value of υ. This has been known for some time, for the reason that a nonparametric estimation method exists for θ_0 in this setup [see Kalbfleisch and Prentice (1980) and Chamberlain (1985)]. In fact, this estimation method is applicable to a model setup that is more general than the MPH model. To proceed, it is useful to distinguish between observed explanatory variables x^* which do not vary within strata, and observed explanatory variables x which do vary within strata. We assume for expositional reasons that the hazard function is multiplicative in a part depending on x^* and a part depending on x. In particular,

$$\theta(t|x^*,x,\upsilon) = \psi(t|x^*,\upsilon) \cdot \theta_0(x), \qquad \upsilon|x^*,x \sim G(\upsilon|x^*,x). \tag{28}$$

This specification allows for full interaction of the values of υ and x^* with the elapsed duration t in the hazard function. This implies that we allow the baseline hazard to differ across strata (i.e., across groups of spells with the same υ). Moreover, υ, x^* and x may be dependent. The basic idea of the estimation method is that a Cox partial likelihood can be constructed *within* strata. For a given stratum, the partial likelihood depends only on θ_0, and not on G or ψ or the values of υ or x^*. These likelihoods can be combined to construct an over-all partial likelihood which can be used to estimate θ_0 (see the above references for details).

Clearly, the effects of the explanatory variables x^* cannot be estimated from this. In other words, to be able to estimate the effect of an observed explanatory variable with this approach, it is essential that the values of the variable sometimes differ across spells within a stratum. In case of two spells per stratum, this amounts to $x_1 \neq x_2$. To see this, note that within such a stratum,

$$\Pr(t_1 > t_2|x_1,x_2,\upsilon) = \frac{\theta_0(x_2)}{\theta_0(x_1) + \theta_0(x_2)},$$

which is only informative on θ_0 if $x_1 \neq x_2$.

The within-stratum baseline hazard ψ as a function of t can subsequently be estimated nonparametrically. Yamaguchi (1986) surveys these methods. Kalbfleisch and Prentice (1980) and Ridder and Tunalı (1999) contain useful expositions on the inclusion of time-varying covariates.

What does this “stratified partial likelihood” estimation approach imply for the identification of θ_0 in the MPH model with multi-spell data? This function is identified up to a multiplicative constant if θ_0, ψ and G in Equation (28) satisfy regularity assumptions corresponding to Assumptions 1–4, and if x varies between spells within strata. Again, we do not need independence of observed and unobserved explanatory variables, and we do not need an assumption on the tail of the distribution of the unobservables. Note that the identification result is valid under a specification of the hazard function that is much more general than the MPH specification.

The approach of the previous paragraphs is particularly appealing if the individual υ are regarded as incidental parameters. With full ML, such parameters can in general not be estimated consistently if asymptotically the number of strata goes to infinity with a fixed number of spells per stratum [Lancaster (2000a)]. In the above approach, however, these parameters cancel out of the partial likelihood. Somewhat loosely one may say that if multiple durations are available for each υ, then duration analysis becomes similar to standard dynamic panel data analysis, where one can get rid of the so-called "fixed effects" before estimating the other parameters. This raises the question to what extent first-differencing of the durations within strata can also be applied to get rid of υ. It seems that this is only feasible if the baseline hazard has a particular functional-form specification, notably the Weibull specification. Assume that the duration dependence is described by $\alpha t^{\alpha-1}$ for all spells and strata. In addition, assume that υ is the same for all spells in a stratum, and assume for convenience that $\theta_0(x_i) = \exp(x_i'\beta)$. For two spells t_1, t_2 within a stratum, with observed explanatory variables x_1 and x_2, respectively, the difference of Equations (26) gives

$$\log t_1 - \log t_2 = -\frac{\beta}{\alpha}(x_1 - x_2) + \frac{\varepsilon_1 - \varepsilon_2}{\alpha}.$$

Note that $\varepsilon_1 - \varepsilon_2$ has a fully specified distribution (as the difference of two i.i.d. EV1 random variables). Thus, with Weibull duration dependence, first-differencing results in an equation from which the Weibull parameter and the systematic hazard can be reliably estimated without the need to make any assumption on the unobserved heterogeneity distribution. Indeed, υ and x are allowed to be dependent.

The identification results discussed in this subsection have been of enormous importance for applied duration analysis. If two observations are available for each υ then the identification of the model does not require an untestable assumption on the tail of the unobserved heterogeneity distribution G anymore, and υ and x need not be independent anymore. We only need some fairly innocuous regularity assumptions and normalizations (of course, in addition to proportionality assumptions on the hazard function). The recent applied literature contains a number of studies showing that the estimates of the parameters of interest are robust with respect to the functional-form specification of G, in case of multiple observed durations for each υ [see Nielsen et al. (1992), Guo and Rodríguez (1992), Gönül and Srinivasan (1993) and Bonnal, Fougère and Sérandon (1997)]. These results are in sharp contrast to those found for the single-spell model (Section 5). It should also be noted that Hahn (1994) finds that his result on singularity of the information matrix in the case of single-spell data (see Subsection 5.6) does not carry over to the case of multi-spell data. Moreover, the stratified partial likelihood estimators are $\sqrt{n}$-consistent.

We finish this section by mentioning an important caveat with multi-spell data. This concerns the fact that the analysis of multi-spell data is particularly sensitive to censoring. With single-spell data, many types of censoring are innocuous in the sense that their effect can be captured by standard adjustments to the likelihood function [see Andersen et al. (1993); recall also the discussion in Subsection 4.2]. With multi-spell data, one has to be more careful. Consider the case where two durations t_1 and t_2 follow each other in time, and where the data are subject to right-censoring at a fixed duration after the common starting point of the t_1 durations. Then the moment at which t_2 is right-censored is not independent from t_2 itself. To see this, consider individuals for which υ is large. For these individuals, t_1 will on average be short. As a result, t_2 will on average start at a relatively early moment. This in turn implies that t_2 will often be right-censored at a relatively high duration. In sum, t_2 and the variable determining the moment at which it is censored are both affected by the unobserved characteristic υ. This violates the standard censoring assumptions of duration analysis [see Visser (1996) for general results, and Keiding (1998)]. As a result, standard partial likelihood estimation methods (like the one above) cannot be applied. Moreover, one cannot estimate (characteristics of) the distribution of t_2 in isolation from t_1 [see Ridder and Tunalı (1999) for an informative exposition]. With censoring in general, first-differencing (like above) is not possible. Finally, the value of t_1 may even affect the probability that the beginning of the second spell is observed at all, in which case a subsample of individuals for which both t_1 and t_2 are observed is selective (this is even true if there is no unobserved heterogeneity)[52]. Of course, with censoring, one may still use standard ML estimation methods with random effects. However, if the realization of t_2 is often unobserved then the use of multi-spell data does not provide much gain over the use of single-spell data. In sum, the less censoring in the data, the larger the advantages of multi-spell data.

7. An informal classification of reduced-form multiple-duration models

In general one may think of many different ways to model a relation between duration variables. In the applied econometric literature on the estimation of multiple-duration models, the range of different models is actually not so large. In this section we provide a rather informal model classification that covers most of the models used in practice[53]. The next sections examine the models in more detail. It should be stressed that we are not concerned with abstract point processes where the durations between events can be related for many reasons [see e.g., Snyder and Miller (1991) for a survey]. Also, we are not concerned with the multiple-duration models in engineering where the lifetime of a system depends on the lifetimes of its components. The latter models are often not very useful to describe economic behavior [although they are an important input in economic analyses of machine maintenance; see e.g., Ryu (1993)]. As we shall see, some of the models that we consider are more natural when dealing with successive spells in a given state or with successive spells in different states[54], whereas others are more natural in the case of competing risks, and yet others are useful in all these

[52] In a recent working paper, Woutersen (2000) develops consistent GMM-type estimators that deal with a number of these problems, while treating unobserved heterogeneity as a fixed effect.

[53] See Hougaard (1987) for an older classification, based on statistical model properties.

[54] Again, what constitutes a state depends on the application at hand (i.e., depends on the relevant underlying theoretical framework). It is possible that what in one application are regarded as multiple

cases. In fact, the recent empirical literature often uses models that simultaneously allow for two different types of dependence of the duration variables. The MPH model with multi-spell data (Section 6) can also be interpreted as a multiple-duration model, as it specifies the joint distribution of the durations in the spells that an individual experiences. We shall see that this specification is in fact a special case of a popular type of multiple-duration model. For expositional reasons we shall restrict ourselves to two duration variables throughout the remainder of this chapter.

"Lagged" durations. The first popular type of dependence concerns an effect of a realized past duration on the current hazard. This type of dependence was introduced by Heckman and Borjas (1980). Suppose that two durations t_1 and t_2 each follow their own PH model, with $\theta_1(t_1|x_1) = \psi_1(t_1)\,\theta_{0,1}(x_1)$ and $\theta_2(t_2|t_1,x_2) = \psi_2(t_2)\,\theta_{0,2}(x_2)\,\xi(t_1)$, where t_2 starts at or after the moment at which t_1 is realized. Basically, this dependence is modeled by including t_1 as an additional covariate in the hazard for t_2. Usually, the underlying economic theory provides a causal interpretation for this type of dependence[55]. Because of the analogy to a regression model with lagged endogenous variables among the explanatory variables, this dependence is sometimes called "lagged-duration dependence". Obviously, different types of restrictions can be imposed on the model determinants $\theta_{0,1}, \theta_{0,2}, \psi_1$, and ψ_2. For example, if t_1 and t_2 denote durations in the same state then it may be imposed that $\psi_1 \equiv \psi_2$, $x_2 = x_1$, and/or $\theta_{0,2}(x_2) = \theta_{0,1}(x_1)$.

Instead of including the value of t_1 in the individual hazard for t_2, one may also use an indicator of whether the individual has been in the state associated with t_1 during the year before the start of t_2, or indeed any other realization of past behavior. In applied labor economics, these types of dependence have been incorporated in reduced-form models for the effects of labor market programs on subsequent unemployment durations and employment durations. It should be stressed however that these studies also allow for other dependencies; see below for examples.

Recently, in financial econometrics, lagged-duration dependence models have been used for the analysis of durations between successive market events such as a buy or sell of a security on a stock market [see e.g., Engle and Russell (1998) and Bauwens and Giot (1998)]. In these models, the hazard function of the ith duration depends on the realizations of previous durations by way of an autoregressive scheme. The baseline hazard is assumed to have a Weibull specification with a single common parameter for all durations.

durations in the same state, are regarded in another application as durations in different states. In practice, for a given individual and a given definition of states, the specifications for the marginal distributions of different spells in a given state are similar, whereas the specifications for the marginal distributions of spells in different states do not contain common parameters or functions.

[55] Here and elsewhere, the relation between the duration variables can be formulated by using the concept of Granger-noncausality. However, for the basic models examined in this chapter, there is no gain from doing this [see Abbring (1998)]. See Florens and Fougère (1996) for a formal analysis of causality in more general continuous-time processes.

Shocks. The second popular type of dependence concerns situations where two durations occur simultaneously, and where the realization of one duration variable has an immediate effect on the hazard of the other duration variable. This type of dependence has been introduced by Freund (1961). To focus the mind, suppose that the realization of t_1 affects the level of the hazard of t_2 afterwards. This can be captured by the inclusion of an indicator of whether t_1 is realized, as a time-varying regressor in the hazard specification of t_2. For example, the hazard of t_2 can be specified as $\psi_2(t_2)\exp(x_2'\beta_2 + \delta \mathrm{I}(t_1 < t_2))$, where $\mathrm{I}(\cdot)$ denotes the indicator function, which is 1 if its argument is true and 0 otherwise. From Subsection 4.2 we know that such a specification requires conditions on t_1. Anticipation by the individual of the future realization of t_1 is ruled out. Note that the individual is allowed to know the (determinants of the) probability distribution of t_1.

The underlying economic theory often provides a causal interpretation for the above type of dependence. Obviously, t_1 and t_2 denote durations in different states, so it does not make sense to impose restrictions across the two hazards.

In practice, it may be too restrictive to assume that the realization of t_1 merely affects the *level* of the hazard of t_2. More generally, the realization may be allowed to affect the whole shape of the hazard of t_2 after the realization of t_1 [56]. In applied econometrics, such types of dependence have been incorporated in reduced-form models for the effect of certain treatments[57] on worker labor-market behavior; we return to this below. In addition, the model described above can be seen as a special case of models in which an individual experiences different stochastic processes which affect each other by way of shifts in the hazard for one process if the other process generates an event. The latter type of models have been used to study the interaction between marital status, number of children, health status, and labor market status. For example, if an unemployed woman marries then her transition rate to employment may drop. It should again be stressed that these studies often also allow for other types of dependence between the duration variables; see below.

Related unobserved determinants. The third type of dependence between duration variables concerns dependence by way of their unobserved determinants. Specifically, consider two durations t_1 and t_2 which each follow their own MPH model, so $\theta_i(t_i|x_i, v_i) = \psi_i(t_i)\,\theta_{0,i}(x_i)\,v_i$, with $i = 1, 2$. Then the dependence between t_1 and t_2 given x is modeled by allowing v_1 and v_2 to be related. In Subsection 8.1 below we provide a more precise definition. This multivariate extension to the MPH model is called the Multivariate Mixed Proportional Hazard (MMPH) model. This has in fact been the

[56] In an empirical analysis of panel survey attrition, Van den Berg, Lindeboom and Ridder (1994) examine a slightly different model in which there is a positive *probability* that t_2 is realized immediately after realization of t_1. Here, t_1 and t_2 are the duration until the individual respondent makes a transition to another labor market state, and the duration until attrition from the panel, respectively.

[57] In biostatistics, θ_0 is often called the treatment effect if x captures whether the subject has received a treatment at the beginning of the spell. Here, we avoid that terminology, and we reserve the term "treatment" for treatments occurring during a spell.

most popular multiple-duration model by far[58]. Note that the relation between the durations is spurious to the extent that it results from the fact that we do not observe v_i.

The MMPH model applies to cases where the two durations occur simultaneously (possibly with the same starting point) as well as to cases where they occur successively. Again, different types of restrictions can be imposed on the model determinants $\theta_{0,1}, \theta_{0,2}, \psi_1, \psi_2$, and the joint distribution $G(v_1, v_2)$, depending on the extent to which t_1 and t_2 represent durations in the same state. Clearly, the MPH model of Section 6 with a single state and multi-spell data is the special case with $\theta_{0,1} = \theta_{0,2}, \psi_1 = \psi_2$, and $v_1 = v_2$.

The MMPH model is regarded as a convenient and flexible model for dependent durations. Of course, there are often good reasons to suspect the presence of important related unobserved determinants, and by now there is an abundant applied literature in which MMPH models are estimated. In the econometric contributions to this literature, the variety of types of states and durations that are considered is vast. Flinn and Heckman (1982b, 1983), Coleman (1990) and Rosholm (1997) estimate MMPH models for the durations of unemployment, employment, etc., in order to study transition rates between different labor market states. Generally, the unobserved determinants of the durations spent in different states are allowed to be related, and the unobserved determinants of different durations spent by an individual in the same state are assumed to be identical. In their studies of attrition in longitudinal panel survey data Van den Berg, Lindeboom and Ridder (1994), Carling and Jacobson (1995) and Van den Berg and Lindeboom (1998) estimate MMPH models for the joint durations of labor-market spells (like a spell of unemployment or a job spell) and the duration of panel survey participation. Lillard and Panis (1998) include attrition in a similar way in their model for the joint durations of marriage, non-marriage, and life. Note that this approach to attrition is in line with the popular modeling setup for sample selection introduced by Heckman (1979).

As we saw in Section 6, MPH models are sometimes estimated under the assumption that the unobserved heterogeneity term is identical across different physical individuals within some group or stratum. Sastry (1997) extends this setup by allowing each individual to belong to two groups with different aggregation levels (families and towns). There is unobserved heterogeneity across each type of group. This effectively amounts to an MMPH specification for the durations of members of different families living in the same town. Similarly, the approach in studies on lifetime durations where the unobserved heterogeneity terms are assumed to be identical across siblings can be generalized to allow v_1 and v_2 for siblings to be a sum of a common determinant and an independent person-specific component (see e.g., Petersen (1996), Yashin and Iachine (1997) and Zahl (1997) for applications][59]. Such a specification for G has gained less

[58] Flinn and Heckman (1982b) provide an early analysis of this model.

[59] The applications of this paragraph illustrate a disadvantage of the “multi-state/multi-spell” terminology: sometimes two spells are in the same state but one does not want to impose that the unobserved heterogeneity terms are identical, so that the multi-spell setup of Section 6 does not apply.

popularity in econometrics for the obvious reason that in econometric applications the association of unobserved heterogeneity to genetic factors is less compelling.

Combinations of dependencies. The presence of related unobserved determinants is particularly important if one is interested in one of the other two types of dependence that we described above. The estimate of the causal effect will be biased if one ignores the spurious dependence that results from the related unobserved determinants. To deal with this, the empirical model should take account of this spurious dependence. The model should allow both for a causal effect *and* for related unobserved heterogeneity.

As examples of a combination of lagged duration dependence and related unobserved heterogeneity, see Heckman, Hotz and Walker (1985), who allow "lagged" durations between the births of previous children to affect the hazard of the duration of the current birth interval, and who allow for correlated unobserved heterogeneity as well [see Omori (1997) and Lancaster (2000b) for other examples]. Lillard (1993), Lillard and Panis (1996), Abbring, Van den Berg and Van Ours (1997), Eberwein, Ham and LaLonde (1997) and Van den Berg, Van der Klaauw and Van Ours (1998) analyze models where the realization of one duration variable has an immediate effect on the hazard of the other duration variable, allowing for related unobserved heterogeneity in order to deal with selectivity. Let us examine them in somewhat more detail. Abbring, Van den Berg and Van Ours (1997) and Van den Berg, Van der Klaauw and Van Ours (1998) study the effect on the exit rate out of unemployment of a punishment for insufficient search effort. The duration until punishment is modeled by way of an MPH model, and the exit rate out of unemployment permanently shifts to another level at the moment the punishment is applied. Lillard (1993) estimates a model for the joint durations of marriage and time until conception of a child, and his model allows the rate at which the marriage dissolves to shift to another level at moments of child birth. Lillard and Panis (1996) estimate a model on the joint durations of marriage, non-marriage, and life, and their model allows the death rate to shift to another level at moments of marriage formation and dissolution. Eberwein, Ham and LaLonde (1997) estimate a (discrete-time) model for the effect of participation in training programs on individual labor market transitions, and they allow the exit rate out of unemployment to shift to another level at the moment of inflow into the program. See Van den Berg, Holm and Van Ours (2001) for a similar analysis in continuous time. In all these applications, we need to rule out anticipations of the realizations of t_1, but the individual is allowed to know the (determinants of the) probability distribution of t_1.

In the applied literature on the effects of training on unemployment durations, "training" is often regarded to be a separate labor market state, and the effect of training on subsequent labor market transitions can then be captured by a model with lagged-duration dependence (or a model where the fact that one has had any training is allowed to affect subsequent transitions). In order to deal with selectivity of those who enrol in training, it is important to allow for related unobserved heterogeneity terms affecting the inflow into training as well as the other transition rates. Gritz (1993) and Bonnal, Fougère and Sérandon (1997) contain sophisticated examples of such analyses. Ham and LaLonde (1996) use experimental data to estimate models for the effects of training on individual labor market transition rates.

In the *absence* of unobserved heterogeneity, the specification, identification, and ML estimation of models with lagged-duration dependence is relatively straightforward. The same holds for models with changes in the hazard of one duration in response to realization of the other duration [given appropriate assumptions on the direction of the causality; see Florens and Fougère (1996)]. However, models with related unobserved heterogeneity terms are less transparent. In the next section we therefore examine MMPH models in detail. Subsequently, in Section 9, we briefly examine the models where related unobserved heterogeneity is combined with a "causal" effect of one duration on the other (that is, we examine a combination of lagged duration dependence and unobserved heterogeneity, and a combination of a shift in the hazard and unobserved heterogeneity).

Some theoretical considerations. We finish this section by stressing that, like in Section 4, it is often not clear to what extent the reduced-form specifications of the dependence between two durations can be justified by economic-theoretical models. This is particularly true for models where the hazard of one duration immediately changes in response to the realization of the other duration. In many cases, individuals may anticipate the realization of the other duration, and the moment at which the anticipation starts is often unobserved. In applications this has to be examined carefully.

In the analysis of MMPH models, as a rule, the assumed parametric family of the joint unobserved heterogeneity distribution $G(v_1, v_2)$ treats v_1 and v_2 in a symmetric way: given the unknown parameters of G, the role of v_1 and v_2 in $G(v_1, v_2)$ can be interchanged without changing G. In particular, if G is continuous then the supports of v_1 and v_2 are assumed to be the same, and if G is discrete then the numbers of points of support are assumed to be the same for v_1 and v_2. It is sometimes difficult to justify such symmetric distributions with economic theory. If, according to the theory, individuals improve their situation when ending one spell and starting another, then the characteristics associated with the second spell should be "superior" in some sense to those of the first spell. If v_1 represents the characteristics of the first spell and v_2 of the second, then this suggests that the support of v_2 should depend on the realization of v_1. Consider for example the on-the-job search model discussed in Subsection 5.5.3. If one observes two consecutive job spells and if the wages are unobserved, then the unobserved heterogeneity term of the second spell exceeds the term of the first spell. Unfortunately, such bivariate heterogeneity distributions have not yet been studied [see Koning et al. (2000) for an application in a structural analysis of an on-the-job search model].

Finally, we address whether the hazards of different durations of the same individual depend on the same set of explanatory variables or not. Economic theory often predicts that both hazards depend on the individual's behavior, and that the forward-looking individual's optimal strategy depends on all structural determinants. For example, in a job search model with two possible employment destination states, the decision on whether to accept a job offer depends on the arrival rates and wage offer distributions of both types of employment, regardless of the employment type of the actual offer [see Thomas (1998)]. In such cases, if the observed explanatory variables are characteristics of the individual himself, then it does not make sense to exclude elements of x from one hazard that are included in the other hazard. In other words, in such cases, $x_1 = x_2$ [note incidentally that this provides an argument against the assumption that unobserved heterogeneity is independent across spells for a given individual; see also Lillard (1993)]. In the event that the researcher observes a determinant of one of the hazards whereas this determinant is assumed to be unobserved by the individual, then it makes sense to include this determinant only in the corresponding hazard. Finally, if one hazard is mechanical and independent of the individual's behavior then obviously it does not need to depend on the determinants of the other hazard [see Van den Berg (1990b) and Ryu (1993) for examples].

8. The Multivariate Mixed Proportional Hazard model

8.1. Definition

In this subsection we define the MMPH model. Next, Subsection 8.2 deals with identification of this model under different situations with respect to the timing of the two underlying spells. We assume that the situation is either such that both durations always start at exactly the same point of time, or that one duration necessarily follows the other. In Subsection 8.3 we discuss parametric specifications for the joint distribution of unobserved heterogeneity and the degree of flexibility of the corresponding models.

For the sake of convenience, we again use the term "individual" to denote the subject that experiences certain spells. In the first situation with respect to the timing of the spells (starting at the same time) we consider the population of individuals in the inflow into the states corresponding to the duration variables, whereas in the second situation (successive durations) we consider the population of individuals in the inflow in the state corresponding to the first duration. Flinn and Heckman (1982b), Chesher and Lancaster (1983) and Ham and LaLonde (1996) consider less "ideal" sampling designs.

We assume that all individual differences in the hazard function of t_1 can be characterized by observed explanatory variables x and unobserved characteristics v_1. Similarly, all individual differences in the hazard function of t_2 can be characterized by observed explanatory variables x and unobserved characteristics v_2. (Of course, one may impose exclusion restrictions on the set of elements of x that is allowed to affect the systematic hazard $\theta_{0,i}(x)$ associated with exit i.) For an individual with explanatory variables x, v_1, v_2, the hazard functions of t_1 and t_2 conditional on x, v_1, v_2 are denoted by $\theta_1(t_1|x, v_1)$ and $\theta_2(t_2|x, v_2)$. The MMPH model is now defined by

Definition 2. *MMPH model*: There are functions ψ_1, ψ_2, $\theta_{0,1}$, $\theta_{0,2}$ such that for every t_1, t_2, x, v_1, v_2 there holds that

$$\theta_1(t_1|x, v_1) = \psi_1(t_1) \cdot \theta_{0,1}(x) \cdot v_1, \quad \theta_2(t_2|x, v_2) = \psi_2(t_2) \cdot \theta_{0,2}(x) \cdot v_2. \tag{29}$$

For convenience, we take ψ_1, ψ_2, $\theta_{0,1}$, $\theta_{0,2}$, v_1, v_2 and the distribution G of v_1, v_2 in the population to satisfy the regularity assumptions that correspond to Assumptions 1–4 for ψ, θ_0, v, G in the MPH model.

Conditional on x, v_1, v_2, the durations t_1 and t_2 are independent. Conditional on x, the variables t_1 and t_2 are only dependent if v_1 and v_2 are dependent. So, in the case of independence of v_1 and v_2, the model reduces to two unrelated ordinary MPH models for t_1 and t_2.

In terms of a regression specification with $\theta_{0,i}(x) = \exp(x'\beta_i)$, this model can be rewritten as

$$\begin{aligned} \log \int_0^{t_1} \psi_1(u)\,\mathrm{d}u &= -x'\beta_1 - \log v_1 + \varepsilon_1, \\ \log \int_0^{t_2} \psi_2(u)\,\mathrm{d}u &= -x'\beta_2 - \log v_2 + \varepsilon_2, \end{aligned} \tag{30}$$

where ε_1 and ε_2 are i.i.d. EV1 distributed, but where v_1 and v_2 may be related.

Now consider the joint distribution of t_1 and t_2 given x. The joint density $f(t_1, t_2|x)$ can be expressed as

$$f(t_1, t_2|x) = \int_0^\infty \int_0^\infty f_1(t_1|x, v_1) f_2(t_2|x, v_2)\,\mathrm{d}G(v_1, v_2),$$

in which we already implicitly assume that v_1, v_2 are independent of x, and in which the probability density function of $t_i|x, v_i$ is for convenience denoted by $f_i(t_i|x, v_i)$. The latter density can of course be expressed in terms of the determinants of θ_i (see Section 2). Let $z_i(t_i)$ denote the integrated baseline hazard associated with t_i. The joint survivor function of t_1 and t_2 can then be expressed as

$$\overline{F}(t_1, t_2|x) = \int_0^\infty \exp(-z_1(t_1)\,\theta_{0,1}(x)\,v_1 - z_2(t_2)\,\theta_{0,2}(x)\,v_2)\,\mathrm{d}G(v_1, v_2).$$

In many applications, the individual likelihood contribution is based on the density above (that is, if the unobserved heterogeneity terms are not treated as incidental parameters). In terms of panel data analysis, this means that v_1, v_2 are treated as "random effects" when estimating the model with Maximum Likelihood.

8.2. Identification results

In this subsection we consider identification results for the MMPH model. It is important to stress that no parametric functional form assumptions are made on the underlying functions $\theta_{0,i}$, ψ_i and G, so, as in Subsection 5.2, we are concerned with *nonparametric* identification.

8.2.1. Competing risks

Recall from Subsection 8.1 that we consider two different situations with respect to the timing of the two spells. In the first situation, both spells start at the same point of time for a given individual, and the individual is observed until the first duration is completed. This is called a competing-risks model, as one may envisage the individual having two options to leave the current state, and the realization of one option is necessary and sufficient for leaving the state. In the second situation with respect to the timing of the spells, the two spells cannot overlap. Moreover, in the second situation both durations can be followed until completion, so there is more information available than in the first situation (see Subsection 8.2.2 below).

In the competing-risks setting, the data provide information on $\min\{t_1, t_2\}$ and on $\arg\min_i t_i$ (i.e., on which duration is the one that ends first). So assume that the data provide the distribution of this "identified minimum". It is well known that this does not suffice to identify the most general competing-risks model (with an arbitrary joint distribution for t_1, t_2, without covariates). In particular, for every model with dependent t_1, t_2 there is an observationally equivalent model with independent t_1, t_2 [see e.g., Lancaster (1990)].

Now let us assume that t_1 and t_2 are generated by an MMPH model with regularity assumptions corresponding to Assumptions 1–4. As in Subsection 5.2, some additional assumptions are needed for identification. These include the equivalents of Assumption 5 (so x is independent of v_1, v_2), Assumption 7 (normalizations), and Assumption 8 ($\mathrm{E}(v_i) < \infty$). In addition, we need to strengthen Assumption 6 on the dispersion of x.

Assumption 9. Variation in observed explanatory variables in the competing-risks setting: *The functions $\theta_{0,1}(x)$, $\theta_{0,2}(x)$ attain all values in a set $(0, \overline{\theta}_{0,1}) \times (0, \overline{\theta}_{0,2})$ with $0 < \overline{\theta}_{0,1}, \overline{\theta}_{0,2}$, when x varies over the set $\mathcal{X}$ of possible values of x.*

If $\theta_{0,i}(x) = \exp(x'\beta_i)$ then sufficient for this is that x has *two* continuous covariates which affect both hazards θ_i but with different coefficients for different i, and which are not perfectly collinear. Moreover, in the population, these covariates must attain all values ranging to minus infinity.

Heckman and Honoré (1989) prove the nonparametric identification of the model under these assumptions. In fact, they strengthen Assumption 9 by taking $\overline{\theta}_{0,i} = \infty$, because they examine a class of models that is somewhat more general than the class of MMPH models [see Abbring and Van den Berg (2000b)]. In any case, note that Assumption 9 is stronger than Assumption 6 on the range of values that θ_0 attains in the MPH model. This is not surprising. However, it is important to note that the identification does not require exclusion restrictions on the hazard specification of either duration. Moreover, identification does not require parametric functional form restrictions on the distribution of unobserved heterogeneity. In the case of binary data on the "identified minimum" (i.e., it is observed which duration ends first but not when) such restrictions are necessary to achieve identification. This illustrates the fact that the timing of events in duration data provides a valuable source of information concerning the underlying model.

It is interesting to obtain some insight into the identification of whether the durations are dependent or not, since this distinguishes the above identification result from the earlier literature in which competing risks models without covariates were examined. In the sequel of this subsection we use T_1, T_2 to denote the random duration variables, and t_1, t_2 to denote realizations of these. We define

$$\theta_1^*(t_1|x, T_2 > t_1),$$

to be the hazard of the duration T_1 at the value t_1, conditional on x and conditional on the duration T_2 exceeding t_1. More generally, the hazard $\theta_1^*(t_1|x, T_2 > t_2)$ corresponds to the conditional distribution of $T_1|x, T_2 > t_2$. We evaluate this hazard for given t_1 and t_2, and in fact we take $t_2 = t_1$. Obviously, the hazard $\theta_2^*(t_2|x, T_1 > t_2)$ can be defined analogically. It is important that the "conditional" hazards $\theta_1^*(t_1|x, T_2 > t_1)$ and $\theta_2^*(t_2|x, T_1 > t_2)$ are observable quantities, as they can be expressed in terms of the distribution of the data. (Note that the "marginal" hazards $\theta_i(t_i|x)$ are unobserved due to the competing risks setting.)

If v_1 and v_2 are independent, then

$$\theta_1^*(t_1|x, T_2 > t_1) = \theta_1(t_1|x) \text{ and } \theta_2^*(t_2|x, T_1 > t_2) = \theta_2(t_2|x).$$

The assumption in Heckman and Honoré (1989) on the values that can be attained by $\theta_{0,i}(x)$ implies that $\theta_{0,1}(x)$ and $\theta_{0,2}(x)$ are not perfectly related, and that there is some independent variation in both. As a result, if v_1 and v_2 are independent then $\theta_{0,2}(x)$ does not affect $\theta_1^*(t_1|x, T_2 > t_1)$, and $\theta_{0,1}(x)$ does not affect $\theta_2^*(t_2|x, T_1 > t_2)$.

Now let us examine what happens if v_1 and v_2 are dependent. It is straightforward to show that

$$\theta_1^*(t_1|x, T_2 > t_1) = \frac{\mathrm{E}_v\left[\theta_1(t_1|x, v_1)\exp\left(-\int_0^{t_1}\theta_1(u|x, v_1)\,\mathrm{d}u - \int_0^{t_1}\theta_2(u|x, v_2)\,\mathrm{d}u\right)\right]}{\mathrm{E}_v\left[\exp\left(-\int_0^{t_1}\theta_1(u|x, v_1)\,\mathrm{d}u - \int_0^{t_1}\theta_2(u|x, v_2)\,\mathrm{d}u\right)\right]},$$

with θ_i as in Equation (29), and with E_v denoting the expectation with respect to the bivariate distribution $G(v_1, v_2)$. If we differentiate this with respect to $\theta_{0,2}(x)$ then the resulting expression has the same sign as

$$-\mathrm{Cov}(v_1, v_2|x, T_1 > t_1, T_2 > t_1)$$

(provided that $t_1 > 0$). If v_1 and v_2 are dependent then in general there are many values of t_1 such that the above expression is nonzero.

In sum, the derivative of $\theta_1^*(t_1|x, T_2 > t_1)$ with respect to $\theta_{0,2}(x)$ and its mirror image for t_2 are informative on the dependence or independence of the unobserved heterogeneity terms. This is intuitively very plausible. If the systematic hazard of t_2 does not directly affect the individual hazard of t_1 but does affect the observed hazard of t_1 then this indicates that there is a spurious relation between the durations by way of their unobserved determinants. It should again be stressed that this is not based on an exclusion restriction in the usual sense of the word. All explanatory variables are allowed to affect (the means of) both duration variables – they are just not allowed to affect the whole duration distributions in the same way[60].

The above results are based on the availability of "single-spell" data. In the present context, this means that for each individual in the sample there is one observation of the "identified minimum" (which consists of $\min\{t_1, t_2\}$ and $\arg\min_i t_i$). Now suppose that the individual-specific value of the υ_1, υ_2 pair is invariant over time. In a recent working paper Abbring and Van den Berg (2000b) show that some of the assumptions made by Heckman and Honoré (1989) can be weakened substantially if the data provide multiple observations on the identified minimum for each individual.

8.2.2. Successive durations

If the two spells are successive, and both durations can be followed until completion, then the data provide the joint distribution $F(t_1, t_2|x)$. In fact, it is merely for expositional reasons that we take the spells to be successive: if they occur (partly) simultaneously and are both observed until completion then the results of this subsection are valid as well, provided that the durations satisfy the model as defined in Subsection 8.1.

The most general model specification does not impose restrictions across the marginal duration distributions, so it allows for $\psi_1 \neq \psi_2, \theta_{0,1} \neq \theta_{0,2}$, and $\upsilon_1 \neq \upsilon_2$. For both marginal hazard functions in this model we make regularity assumptions corresponding to Assumptions 1–4. In addition, we adopt the equivalents of the Assumptions 5–8 that were made to identify the MPH model. Honoré (1993) shows that under these assumptions the MMPH model is identified. (Assumptions 6 and 8 may be jointly replaced by Assumptions 6b and 8b.)

This result is not surprising, because the data on $t_i|x$ identify the determinants of the MPH model for t_i (which are ψ_i, $\theta_{0,i}$ and the marginal distribution of υ_i), provided that the assumptions for identification of this MPH model are satisfied. The relation between υ_1 and υ_2 is subsequently identified from the observed relation between t_1 and t_2 given x.

Sometimes it makes sense to impose a priori restrictions across the marginal duration distributions. The most restrictive specification imposes that $\psi_1 = \psi_2, \theta_{0,1} = \theta_{0,2}$,

[60] Of course, the $\theta_{0,i}(x)$ are not directly observed. Heckman and Honoré (1989) identify these by examining data at zero durations. Whether this can be used to construct a useful test statistic on independence remains to be seen.

and $v_1 = v_2$. We already know from Section 6 that this model is identified under weak assumptions. Now let us consider an intermediate case in which we impose that $v_1 = v_2$ but allow the baseline hazards ψ_1 and ψ_2 to be different. In addition, we do not assume that there are observed explanatory variables x. In other words, the analysis is conditional on a given value of x, and we allow for full interaction of the actual value of x with the model determinants: ψ_i may depend on x in an unspecified way, and v and x may be dependent in the population (from this point of view we do not consider an "intermediate" case, as this generalizes the MMPH specification). Thus,

$$\theta_i(t|x, v) = \psi_i(t|x) \cdot v, \qquad v|x \sim G(v|x).$$

This includes of course as a special case that $\psi_i(t|x)$ can be written as $\psi_i(t)\, \theta_{0,i}(x)$. We make regularity assumptions corresponding to Assumptions 2–4. Honoré (1993) shows that this model is identified, provided that a normalization is imposed on the integrated baseline hazard (analogical to Assumption 7). Note that we do not need to make assumptions corresponding to the previously made Assumptions 5, 6 and 8. Perhaps the most important issue here is that identification does not require independence of v and x. In many applications, such independence is difficult to justify. Like in Section 6, if unobserved heterogeneity values are identical across different durations then the model is similar to a standard dynamic panel data model.

8.3. Specification of the bivariate unobserved heterogeneity distribution

8.3.1. Dimensionality

The types of justifications used for parametric functional forms of G in MPH models are often unavailable for MMPH models. This is particularly true for the choice of a specification for the dependence of v_1 and v_2. In this subsection we focus on the choice of the dimensionality of the distribution of G (or more accurately, the dimension of the support of G). In Subsection 8.3.2 we then examine the types of dependence that can be generated by different parametric functional forms for a G with a given dimensionality.

The so-called "one-factor loading specification" has been a popular specification for a bivariate distribution of unobserved heterogeneity terms in MMPH models [see Flinn and Heckman (1982b, 1983) for early applications, and Heckman, Hotz and Walker (1985), Heckman and Walker (1987, 1990) and Bonnal, Fougère and Sérandon (1997) for subsequent applications]. This specification reduces the dimensionality of the distribution G from 2 to 1. In particular, it assumes that there is a univariate random variable z such that

$$v_i = \exp(\alpha_i + \gamma_i z) \qquad i = 1, 2. \tag{31}$$

(Note that this z does not refer to the integrated baseline hazard here.) This specification can be straightforwardly generalized to a higher number of different durations as well as a higher dimension of the random variable z. If z is two-dimensional then we obtain a "two-factor loading specification", etc.

The two (related) advantages of the "factor loading specifications" are (1) they restrict the number of unknown parameters, leading to a sparse specification, and (2) they limit the computational burden of the estimation of the model. The number of parameters related to G equals the number of parameters of the distribution of z, plus the number of α_i and γ_i parameters, minus normalizations. This typically increases linearly with the number of different durations n. If $\upsilon_1, \dots, \upsilon_n$ has a genuine multivariate distribution then the number of parameters related to G typically increases quadratically with n. To illustrate the computational advantage, consider the case where $\log \upsilon_1, \dots, \log \upsilon_n$ has a multivariate normal distribution. The evaluation of the joint density function of $t_1, \dots, t_n$ then requires the evaluation of an n-dimensional integral. However, if the υ_i are related by a one-factor loading specification then the integral is one-dimensional. See for example Bonnal, Fougère and Sérandon (1997), where $n = 8$. Note that computational burden is less of a problem in the case of discrete υ_i and n smaller than, say, 4.

Hougaard (1987) stresses that it is too restrictive to assume that $\upsilon_1 \equiv \upsilon_2$ if the corresponding spells do not concern the same state. If (i) $\upsilon_1 \equiv \upsilon_2$, and (ii) both durations are always observed, and (iii) each duration is described by an identified MPH model, then the full unobserved heterogeneity distribution is completely identified from data on only one of the durations. We now show that somewhat similar problems may arise in the case of a one-factor loading specification for G.

Indeed, the main disadvantage of the one-factor loading specification concerns the relation it imposes on the marginal duration distributions on the one hand, and the dependence of the durations on the other. If $\mathrm{Var}(\upsilon_1) > 0$ and $\mathrm{Var}(\upsilon_2) > 0$ then it automatically follows that $\mathrm{Cov}(\upsilon_1, \upsilon_2) \neq 0$. So if the data provide evidence for unobserved heterogeneity in the marginal distributions of t_1 and t_2, then the model implies that these durations must be dependent. Similarly, if the durations are independent, then the model implies that there is no unobserved heterogeneity for at least one of the durations. If the dependence between the durations changes, then necessarily the marginal duration distributions change as well. Lindeboom and Van den Berg (1994) show in detail that these may amount to serious restrictions on the specification of the full model.

To illustrate this issue, suppose that the distribution of z belongs to a parametric family of distributions with two parameters: a location parameter μ and a scale parameter σ (for example, z has a normal distribution with parameters μ and σ). Then

$$z = \mu + \sigma \tilde{z},$$

where $\tilde{z}$ has a completely specified distribution. By substituting this into Equation (31), it is clear that we can only identify $\alpha_1 + \gamma_1 \mu$, $\alpha_2 + \gamma_2 \mu$, $\gamma_1 \sigma$ and $\gamma_2 \sigma$. This implies that in effect we only have two parameters at our disposal to capture the 3 second moments of $\log \upsilon_1$, $\log \upsilon_2$ (which are $\mathrm{Var}(\log \upsilon_1)$, $\mathrm{Var}(\log \upsilon_2)$ and $\mathrm{Cov}(\log \upsilon_1, \log \upsilon_2)$).

8.3.2. The dependence between the durations

In this subsection we examine the dependence of the two duration variables in the MMPH model. For this purpose we use some summary measures of the association between two random variables. For a given association measure we focus on two issues: first, which range of values of this association measure can be attained by the MMPH model in general, and secondly, to what extent is this range further narrowed if G is assumed to belong to specific families of distributions. The first issue is of importance for a comparison of the MMPH model to other models for the dependence between duration variables. The second issue is of importance for a comparison of the flexibility of different families of heterogeneity distributions, and to obtain insight into the range of bivariate models that can be generated by a specific G. The results in this subsection are from Van den Berg (1997).

The regression-type specification of the MMPH model (see Equation 30) suggests that $\mathrm{Corr}(\log z_1(t_1), \log z_2(t_2)|x)$ may be an interesting summary measure of the association between t_1 and t_2. Unfortunately it turns out that for our purposes it is not, because it can attain every value in $(-1, 1)$ for given baseline hazards, by choosing an appropriate G. Moreover, it can attain every value in $(-1, 1)$ within the popular parametric families of distributions for G. Consider instead $\mathrm{Corr}(t_1, t_2|x)$, and assume for the moment that the baseline hazards are constant. The correlation of the duration variables is informative on the strength of the linear relationship between these variables. It is a commonly used measure that is readily understood. Here, it equals

$$\mathrm{Corr}(t_1, t_2|x) = \frac{\mathrm{Cov}(\frac{1}{v_1}, \frac{1}{v_2})}{\prod_{i=1}^{2}\left[\mathrm{Var}(\frac{1}{v_i}) + \mathrm{E}(\frac{1}{v_i^2})\right]^{1/2}}. \tag{32}$$

Note that it does not depend on x and that its sign equals the sign of $\mathrm{Corr}(1/v_1, 1/v_2)$.

Van den Berg (1997) shows that

$$-\tfrac{1}{3} < \mathrm{Corr}(t_1, t_2|x) < \tfrac{1}{2},$$

regardless of the values of $\theta_{0,1}(x)$ and $\theta_{0,2}(x)$, and regardless of the shape of $G(v_1, v_2)$ (but provided that the right-hand side of Equation (32) exists). The inequalities are sharp in the sense that they can be approached arbitrarily closely by choosing appropriate G.

The result above (and most of the results below) can be easily generalized to models with Weibull baseline hazards. In that case, the upper and lower bound depend on the parameters of the baseline hazard, but they are always strictly between −1 and 1, and the lower bound is always closer to zero than the upper bound[61].

[61] Similar results can be derived for bivariate accelerated failure time models and bivariate duration models in discrete time, notably the discretized (i.e., rounded-off) bivariate MPH model and the rather popular bivariate discrete-time duration model in which the exit probabilities have logistic specifications.

In the empirical literature, the most frequently used families of distributions for υ_1, υ_2 are (1) the family of bivariate discrete distributions with two points of support for υ_1 and for υ_2, and (2) the family of bivariate normal distributions for $\log \upsilon_1, \log \upsilon_2$. These families include as special cases the one-dimensional distributions with perfect correlations (these can be represented by the one-factor loading specification 31). Coleman (1990), Van den Berg, Lindeboom and Ridder (1994), Carling and Jacobson (1995), and Van den Berg and Lindeboom (1998) adopt multivariate discrete distributions for G[62], whereas Butler, Anderson and Burkhauser (1986), Lillard (1993), Xue and Brookmeyer (1996), Lillard and Panis (1996, 1998) and Ng and Cook (1997) adopt multivariate normal distributions[63]. It turns out that in the discrete case, every value in $(-\frac{1}{3}, \frac{1}{2})$ can be attained. By implication, this is also true in the case of more than two points of support for each υ_i. In the normal case, $\mathrm{Corr}(t_1, t_2|x)$ can only attain values in $[-3 + 2\sqrt{2}, \frac{1}{2})$, where the lower bound equals about -0.17.

The lower bound $-\frac{1}{3}$ is attained for a discrete distribution for υ_1, υ_2 such that $\Pr(\upsilon_1 = c_1, \upsilon_2 = \infty) = \Pr(\upsilon_1 = \infty, \upsilon_2 = c_2) = \frac{1}{2}$, with $0 < c_1, c_2 < \infty$[64]. In that case, the bivariate distribution of $t_1, t_2|x$ is such that, with probability $\frac{1}{2}$, $t_1|x$ is zero and $t_2|x$ has an exponential distribution, and with probability $\frac{1}{2}$ this holds with t_1 and t_2 interchanged. We conclude that in an MMPH model these (and similar) duration distributions cannot be generated if $\log \upsilon_1, \log \upsilon_2$ has a normal distribution, which may be a disadvantage of the latter if one is interested in a flexible specification[65].

For the general model as well as within the parametric families discussed above, the distributions that give the largest and smallest possible value of $\mathrm{Corr}(t_1, t_2|x)$ are such that $\log \upsilon_1$ and $\log \upsilon_2$ are perfectly correlated. This means that the range of values for $\mathrm{Corr}(t_1, t_2|x)$ is the same as in the case of a one-factor loading model (see Equation 31)

[62] Engberg, Gottschalk and Wolf (1990) estimate a bivariate discrete-time duration model in which the individual per-period exit probabilities are logistic functions of $\psi_i(t_i)\,\theta_{0,i}(x)\,\upsilon_i$, and in which G has a bivariate discrete distribution. Meghir and Whitehouse (1997) estimate a similar discrete-time model, with a genuine bivariate discrete distribution, but with probit specifications for the exit probabilities. Heckman, Hotz and Walker (1985), Heckman and Walker (1987, 1990) and Gritz (1993) adopt discrete distributions for z in a one-factor loading specification. Card and Sullivan (1988), Mroz and Weir (1990), Ham and LaLonde (1996) and Eberwein, Ham and LaLonde (1997) estimate discrete-time bivariate duration models with logistic probabilities and a one-factor loading specification for z with a discrete distribution.

[63] Flinn and Heckman (1982b, 1983) and Bonnal, Fougère and Sérandon (1997) adopt normal distributions for z in a one-factor loading specification. In a sensitivity analysis, the latter study also adopts a discrete distribution for z.

[64] This should not be interpreted as an advantage of discrete random variables for υ_1, υ_2 vis-à-vis continuous random variables, for one can construct families of bimodal continuous distributions for G such that $-\frac{1}{3}$ can be approached arbitrarily closely.

[65] Butler, Anderson and Burkhauser (1989) assume υ_1, υ_2 to have a bivariate discrete distribution with points of support that are fixed in advance. This means that the only parameters of G to be estimated are the probabilities associated with these points of support. This can be shown to narrow the range of values of $\mathrm{Corr}(t_1, t_2|x)$ as well, in particular if the points for υ_1 or υ_2 are chosen to be relatively close to one another [see Van den Berg (1997) for examples].

with an appropriate distribution of z. In other words, a reduction of the class of G to one-factor loading specifications does not further restrict the range of values that $\mathrm{Corr}(t_1, t_2|x)$ can attain[66]. From this point of view, one-dimensional random variation in the unobserved heterogeneity terms is sufficient for maximum flexibility in terms of the correlation of the durations.

As an alternative measure of association, consider Kendall's τ (or "Kendall's coefficient of concordance"). This is the most popular global ordinal measure of association in the literature on multivariate durations [see e.g., Genest and MacKay (1986), Oakes (1989) and Guo and Rodríguez (1992)]. There are several equivalent ways to formally define it. The definition given by Kendall (1962) is particularly useful for general multivariate duration models,

$$\tau(t_1, t_2|x) = 4\mathrm{E}(F(t_1, t_2|x)) - 1,$$

where the expectation is taken with respect to $F(t_1, t_2|x)$ itself. Kendall's τ only attains values in $[-1, 1]$. It is an ordinal measure, and it is informative on the strength of any monotone relation. It equals 1 (-1) if and only if t_2 is a monotone increasing (decreasing) function of t_1. Because it is invariant under monotone transformations of the random variables, the value of $\tau(t_1, t_2|x)$ in the MMPH model does not depend of the baseline hazards or on the values of the systematic hazards (so the baseline hazards can be taken as constants, and the conditioning on x can be omitted). As a result, it only depends on the distribution G of the unobserved heterogeneity terms, which is exactly the part of the model that causes the dependence of the durations.

For convenience, assume that $G(v_1, v_2)$ follows a one-factor loading specification, i.e., suppose Equation (31) holds. It turns out that all values between -1 and 1 can be attained by $\tau(t_1, t_2)$, within any family of continuous distributions for z. However, if z (and therefore v_i) is restricted to have a discrete distribution with n points of support ($n = 2, 3, \ldots, \infty$), then

$$-1 + \tfrac{1}{n} < \tau(t_1, t_2) < 1 - \tfrac{1}{n}.$$

These inequalities are sharp in the sense that they are approached arbitrarily closely for appropriate values of the parameters in the one-factor loading specification (31).

The results for τ are clearly quite different from those for the correlation coefficient. This is because τ detects linear and nonlinear monotone relations alike, and it does not depend on the relative magnitudes of the duration variables, but only on their ordering. The fact that the range of values of $\tau(t_1, t_2)$ is restricted for discrete distributions with finite n can be explained as follows. In this case, the population can be subdivided into a finite number of groups of individuals, and

[66] Note that if $v_1 \equiv v_2$ then this range reduces to $(0, 1/2)$.

within these groups, all individuals are the same in terms of their v_1 and v_2. This implies that there is a positive probability that two random drawings of t_1 and t_2 are from the same group. Now consider all observations for a single group. Because they all have the same v_1 and v_2, there is no relation at all between t_1 and t_2 within the group. This restricts the population value of $\tau(t_1, t_2)$. It does not affect the range of values of $\text{Corr}(t_1, t_2|x)$ because the "within-group" lack of correlation can be made quantitatively unimportant by making the "between-group" differences large.

In all cases, the bounds for $\tau(t_1, t_2)$ are attained by "spreading out" the heterogeneity distribution as much as possible. If z is continuous then the resulting bivariate distribution of $t_1, t_2|x$ is such that all probability mass is on a single curve for t_1 and t_2. We conclude that in an MMPH model such a duration distribution cannot be generated if z has a discrete distribution with a finite number of points of support. This suggests that it is useful in empirical applications to try to increase the number of mass points.

We finish this subsection by noting that in applications it may also be interesting to examine the dependence of the residual duration variables if one conditions on survival up to a certain duration. It may also be interesting to examine how the (non-causal) effect of the realization of one duration variable on the hazard rate of the other changes with the realized value of the first duration variable. Oakes (1989), Anderson et al. (1992), Hougaard, Harvald and Holm (1992b) and Yashin and Iachine (1999) provide analyses for the general case, and they also discuss how the dependence patterns are affected by the functional form of G.

9. Causal duration effects and selectivity

9.1. Lagged endogenous durations

In this subsection we briefly examine bivariate duration models with lagged-duration dependence as well as mutually related unobserved heterogeneity terms. Recall from Section 7 that such models have been used to study the impact of the length of an unemployment spell on the length of the next unemployment spell. Also recall that the estimate of the effect of the previous duration is biased if one ignores the spurious dependence from related unobserved determinants.

In terms of the hazards, the model specification reads

$$\begin{aligned}\theta_1(t_1|x, v_1) &= \psi_1(t_1) \cdot \theta_{0,1}(x) \cdot v_1,\\ \theta_2(t_2|t_1, x, v_2) &= \psi_2(t_2) \cdot \theta_{0,2}(x) \cdot \xi(t_1) \cdot v_2,\end{aligned} \tag{33}$$

and we make the following regularity assumption on the function ξ:

Assumption 10: *The function $\xi(t)$ is positive for every $t \in [0, \infty)$.*

If v_1 and v_2 are independent, then, conditional on x, the durations t_1 and t_2 are only dependent if $\xi(t_1)$ is not a constant. In the general case, the joint density of t_1 and t_2 given x is straightforwardly expressed as

$$f(t_1, t_2|x) = \int_0^\infty \int_0^\infty f_1(t_1|x, v_1) f_2(t_2|t_1, x, v_2)\, \mathrm{d}G(v_1, v_2),$$

in obvious notation. Note that if one allows for more than two consecutive spells then in practice there may be initial-conditions problems, as one may not observe the duration of the first spell.

If both durations can be followed until completion, then the data provide the joint distribution $F(t_1, t_2|x)$. Honoré (1993) shows that this model is identified from these data, under some conditions. For both marginal hazard functions in this model we make regularity assumptions corresponding to Assumptions 1–4, and we adopt regularity Assumption 10. In addition, we adopt the equivalents of Assumptions 5, 6b and 7 on v_i, $\theta_{0,i}$ and ψ_i[67]. We also normalize the function ξ, and we replace the equivalent of Assumption 8 by a slightly different assumption:

Assumption 11. Normalization: *For some a priori chosen* t_0*, it holds that* $\xi(t_0) = 1$.

Assumption 12. Tails of the joint unobserved heterogeneity distribution: $\mathrm{E}(v_1) < \infty$ *and* $\mathrm{E}(v_1 v_2) < \infty$.

Sufficient for Assumption 12 is that $\mathrm{E}(v_i^2) < \infty$ for $i = 1, 2$. In sum, we adopt Assumptions 1–4, the equivalents of Assumptions 5, 6b and 7, and Assumptions 10–12.

Here, as in the model with successive durations and $v_1 \neq v_2$ (Subsection 8.2.2), identification requires assumptions on the tails of the distributions of v_1 and v_2 (notably, finiteness of moments), and it requires that the individual hazards are proportional in t and x. It is plausible that these assumptions can be substantially weakened if the data provide multiple observations on t_1, t_2 for each v_1, v_2 pair [see Woutersen (2000) for results].

9.2. Endogenous shocks

In this subsection we examine bivariate duration models with the property that the hazard of the duration t_2 moves to another level at the moment at which the other duration t_1 is completed, with mutually related unobserved heterogeneity terms. Recall from Section 7 that such models have been used to study the effect of punishments and training on the exit rate out of unemployment and the effect of marriage dissolution on the death rate. Also recall that the estimate of the change of the hazard is biased if one ignores the spurious dependence from related unobserved determinants. Finally,

[67] In fact, the differentiability condition in Assumption 6b can be weakened to continuity here.

recall that we need to rule out anticipations of the realizations of t_1, but the individual is allowed to know the (determinants of the) probability distribution of t_1.

We adopt a framework where the two durations start at the same point of time, and where the realization of t_1 affects the shape of the hazard of t_2 from t_1 onwards. The data provide observations of t_2 and x. If t_1 is completed before t_2 then we also observe t_1; if not then we merely observe that t_1 exceeds t_2. The model and data are thus distinctly asymmetric in the two durations. Somewhat loosely, one may say that t_2 is the "main" duration, or the "endogenous duration of interest", whereas t_1 is an "explanatory" duration, and the causal effect of t_1 on t_2 is the "treatment effect".

In terms of the hazards, the model specification reads

$$\begin{aligned}\theta_1(t_1|x,\upsilon_1) &= \psi_1(t_1)\cdot\theta_{0,1}(x)\cdot\upsilon_1,\\ \theta_2(t_2|t_1,x,\upsilon_2) &= \psi_2(t_2)\cdot\theta_{0,2}(x)\cdot e^{\delta \mathrm{I}(t_1<t_2)}\cdot\upsilon_2,\end{aligned}\tag{34}$$

where $\mathrm{I}(\cdot)$ denotes the indicator function, which is 1 if its argument is true and 0 otherwise. If υ_1 and υ_2 are independent, then, conditional on x, the durations t_1 and t_2 are only dependent if $\delta \neq 1$. In the general case, the joint density of t_1 and t_2 given x is straightforwardly derived as in the previous subsection.

In a recent working paper Abbring and Van den Berg (2000a) provide identification results for this model. In fact, they allow δ to depend on past observables. These results are similar to those for Subsection 9.1 in that they require independence of x from υ_1,υ_2, and they require an assumption on the first moments of υ_1,υ_2. If multiple observations are available for each υ_1,υ_2 pair then such assumptions are not needed.

Contrary to models of binary treatments and binary outcomes, the treatment effect δ is identified without the need to rely on exclusion restrictions or parametric functional-form assumptions regarding the distribution of υ_1,υ_2. In particular, the set of explanatory variables affecting $\theta_{0,1}$ does not have to be larger than the set affecting $\theta_{0,2}$, and the joint distribution of υ_1,υ_2 can be any member of a broad nonparametric class of distributions. These results imply that the timing of events conveys useful information on the treatment effect. This information is discarded in a binary framework. In conclusion, duration analysis is useful for the study of treatment effects in non-experimental settings[68, 69].

10. Conclusions and recommendations

Since the early 1980s the econometric analysis of duration variables has become widespread. This chapter has provided an overview of duration analysis, with an

[68] The model of this subsection does not allow the size of the treatment effect to depend on unobserved heterogeneity. Given the recent interest in heterogeneity of treatment effects [see e.g., Heckman, LaLonde and Smith (1999)], it is a challenge for future research to incorporate this into duration analysis. See Abbring and Van den Berg (2000a) for results on this.

[69] Robins (1998) analyzes treatment effects in a different type of duration models where unobserved determinants of the duration of interest may vary over time and may depend on the treatment.

emphasis on the specification and identification of duration models, and with special attention to models for multiple durations.

We have seen that the hazard function of the duration distribution is the focal point and basic building block of econometric duration models. Properties of the duration distribution are generally discussed in terms of properties of the hazard function. The individual hazard function and the way it depends on its determinants are the "parameters of interest". This approach is dictated by economic theory. Theories that aim at explaining durations focus on the rate at which the subject leaves the state at a certain duration given that the subject has not done so yet. In particular, they explain this exit rate in terms of external conditions at the point of time corresponding to that duration and in terms of the underlying economic behavior of the subject given that he is still in the state at that duration.

The Mixed Proportional Hazard model and its special cases are by far the most popular duration models based on a specification of the hazard function. We have seen that the recent mathematical-statistical literature on counting processes has formulated conditions under which time-varying explanatory variables can be included in MPH models in such a way that one can still perform valid econometric inference with standard methods.

The MPH model and its special cases are often regarded to be useful reduced-form models for duration analysis. The resulting estimates are then interpreted with the help of some economic theory. Unfortunately, the proportionality assumption of the (M)PH model can in general not be justified on economic-theoretical grounds. However, if the optimal strategy of the individual is myopic (e.g., because of repeated search, or because the discount rate is infinite), then this proportionality can often be deduced from economic theory.

The MPH model is nonparametrically identified from single-spell data, given an assumption on the tail of the unobserved heterogeneity distribution, like finiteness of its mean. However, the model is nonparametrically unidentified if such an assumption is dropped. Moreover, the adoption of a model that is observationally equivalent to (but different from) the true model leads to incorrect inference on the parameters of interest. This is bad news, as it is often difficult to make any justified assumption on the tail of the unobserved heterogeneity distribution. In applications where the unobserved heterogeneity term represents an important economic variable, economic theory might provide a justification of the finite mean assumption.

Let the finite mean assumption be satisfied. The observed hazard function of the duration given the observed explanatory variables is nonproportional, meaning that it cannot be expressed as a product of a term depending only on the elapsed duration and a term depending only on the observed explanatory variables. With single-spell data, the unobserved heterogeneity distribution in MPH models is identified from the interaction between the duration and the explanatory variables in the observed hazard, or, in other words, from the observed type of nonproportionality of the observed hazard. However, unobserved heterogeneity can not generate just any type of interaction. The class of models for the observed hazard that is generated by

MPH models is smaller than the general class of interaction models for the observed hazard. In other words, the MPH model is overidentified with single-spell data.

In MPH models, the sign of the interaction between the duration and the explanatory variables in the observed hazard is affected by the type of unobserved heterogeneity distribution. However, under weak conditions, the sign is always negative at small durations regardless of the type of heterogeneity distribution. If unobserved heterogeneity has a Gamma distribution, then the interaction is negative at all durations and all values of the systematic part of the hazard function. If unobserved heterogeneity has a discrete distribution with two positive mass points then the interaction is negative at small durations and positive at large durations.

In MPH models, the effect of an explanatory variable on the observed hazard can be negative at some durations even if the explanatory variable has a positive effect on the underlying individual (or systematic) hazard. This means that it is not possible to deduce the sign of the effect of the explanatory variable on the underlying individual hazard from the observed effect of the variable on the observed hazard at certain durations. Fortunately, this remarkable effect can only occur for some local duration intervals.

By now, there is overwhelming evidence that with single-spell data, minor changes in the assumed parametric specification of the MPH model, while leading to a similar over-all fit, may produce very different parameter estimates. Also, very different models may generate similar data. Estimation results from single-spell data are sensitive to misspecification of the functional forms associated with the model determinants, and this sensitivity is stronger than usual in econometrics. In the absence of strong prior information on the model determinants, single-spell data do not enable a robust assessment of the relative importance of these determinants as explanations of random variation in the observed durations. Therefore, interpretations based on estimation results are often unstable and should be performed with extreme caution.

In biostatistics, this state of affairs has led to a renewed interest in Accelerated Failure Time models as alternative reduced-form duration models for the analysis of single-spell duration data. From an econometric point of view, the AFT approach is unsatisfactory, because it does not focus on the parameters of the individual hazard as the parameters of interest. However, if one is only interested in the sign or significance of a covariate effect on the individual durations then the AFT approach may be useful.

In practice, it may be useful to exploit predictions from the underlying economic theory when specifying the duration model, by imposing these as restrictions on the functional form of the heterogeneity distribution or the baseline hazard. It may be even more useful to look for data with multiple spells (see below). Now suppose that these options are not available. Concerning the baseline hazard, the conceived wisdom is that a piecewise constant specification is then the most useful. Such a specification is flexible and convenient from a computational point of view. Concerning the unobserved heterogeneity distribution, it may be useful to start off with an informal examination of the sign of the interaction in the observed hazard. If it is negative at all durations then a Gamma distribution may give a better fit whereas if it is positive at large durations then a discrete distribution may give a better fit.

By now, the empirical analysis of MPH models with multi-spell duration data is widespread. Basically, if two observations are available for each unobserved heterogeneity value, then the identification of the model does not require an untestable assumption on the tail of the unobserved heterogeneity distribution anymore, and, perhaps even more importantly, observed and unobserved explanatory variables are allowed to be dependent. The identification of this distribution does not come anymore from the interaction between the duration and the observable explanatory variables in the observed hazard. Data on multiple spells for the same individual therefore remove the identification problems associated with single-spell data. Moreover, a consensus has emerged that multi-spell data allow for reliable inference that is robust with respect to the specification of the unobserved heterogeneity distribution. Multi-spell duration data make duration analysis more similar to dynamic panel data analysis. It should however be stressed that the analysis of multi-spell data is particularly sensitive to censoring.

The chapter pays special attention to models for multiple durations. Here, the marginal duration distributions need not be the same. In general one may think of many different ways to model a relation between duration variables. In the applied econometric literature on the estimation of multiple-duration models, the range of different models is actually not so large. Typically, the models allow for dependence between the duration variables by way of their unobserved determinants, with each single duration following its own MPH model. In addition to this, the model may allow for a “causal” effect of one duration on the other, as motivated by an underlying economic theory. The first popular type of causal effect concerns an effect of a realized past duration on the current hazard. Basically, this is modeled by including the realized past duration as an additional covariate in the hazard for the current duration. The second popular type of causal effect concerns situations where two durations occur simultaneously, and where the realization of one duration variable has an immediate effect on the hazard of the other duration variable. This includes models of treatment effects in the presence of selectivity and in the absence of exclusion restrictions.

For such models, identification results have been derived which are similar in contents to those for MPH models with single-spell data. The identification conditions can be weakened substantially if multiple observations are available for each value of the heterogeneity pair, or if cross-restrictions are imposed on the distributions of the two durations in the multiple duration model.

The multiple-duration model where the marginal duration distributions each satisfy an MPH specification, and the durations can only be dependent by way of their unobserved determinants, is called the Multivariate Mixed Proportional Hazard (MMPH) model. In the empirical analysis with such models it is important to assume a genuine multivariate distribution for the unobserved heterogeneity terms. Here, “genuine” means that there is no deterministic relation between any two heterogeneity terms. More restrictive specifications, like the one-factor loading specification, impose cross-restrictions on the marginal duration distributions and the dependence of the durations. In such cases, if the data provide evidence for unobserved heterogeneity in the marginal duration distributions, then the model implies that these durations must be dependent. Similarly, in such cases, if the durations are independent, then the model implies that there is no unobserved heterogeneity for at least one of the durations.

Factor loading specifications have been popular because they restrict the number of unknown parameters, leading to a sparse specification, and they limit the computational burden of the estimation of the model. However, the latter can also be achieved by adopting a (multidimensional) discrete distribution for the unobserved heterogeneity terms. In fact, discrete heterogeneity distributions are particularly flexible, in the sense that they are able to generate a relatively wide range of values for the association measures of the corresponding durations. In empirical applications with MMPH models, it is therefore useful for computational reasons and for reasons of flexibility to assume a multidimensional discrete distribution for the unobserved heterogeneity terms. One may then try to increase the number of mass points. If the number of duration types is relatively large then one may reduce the number of parameters of the multidimensional discrete distribution somewhat by imposing, say, a two-factor loading structure.

References

Abbring, J.H. (1998), “Treatment effects in duration models and Granger non-causality”, Working paper (Free University, Amsterdam).

Abbring, J.H., and G.J. Van den Berg (2000a), “The nonparametric identification of treatment effects in duration models”, Working paper (Free University, Amsterdam).

Abbring, J.H., and G.J. Van den Berg (2000b), “The non-parametric identification of the mixed-proportional-hazard competing-risks model”, Working paper (Free University, Amsterdam).

Abbring, J.H., and G.J. Van den Berg (2001), “The unobserved heterogeneity distribution in duration analysis”, Working paper (Free University, Amsterdam).

Abbring, J.H., G.J. Van den Berg and J.C. Van Ours (1997), “The effect of unemployment insurance sanctions on the transition rate from unemployment to employment”, Working paper (Tinbergen Institute, Amsterdam).

Andersen, P.K., and Ø. Borgan (1985), “Counting process models for life history data: a review (with discussion)”, Scandinavian Journal of Statistics 12:97–158.

Andersen, P.K., Ø. Borgan, R.D. Gill and N. Keiding (1993), Statistical Models Based on Counting Processes (Springer, New York).

Andersen, P.K., M.W. Bentzon and J.P. Klein (1996), “Estimating the survival function in the proportional hazards regression model: a study of the small sample size properties”, Scandinavian Journal of Statistics 23:1–12.

Anderson, J.E., T.A. Louis, N.V. Holm and B. Harvald (1992), “Time-dependent association measures for bivariate survival distributions”, Journal of the American Statistical Association 87:641–650.

Anti Nilsen, Ø., and F. Schiantarelli (1998), “Zeroes and lumps in investment: empirical evidence on irreversibilities and non-convexities”, Working paper (University of Bergen, Bergen).

Antonides, G. (1988), Scrapping a Durable Consumption Good (Erasmus Universiteit, Rotterdam).

Arjas, E. (1989), "Survival models and martingale dynamics (with discussion)", Scandinavian Journal of Statistics 16:177–225.

Arroyo, C.R., and J. Zhang (1997), "Dynamic microeconomic models of fertility choice: a survey", Journal of Population Economics 10:23–65.

Baker, M., and A. Melino (2000), "Duration dependence and nonparametric heterogeneity: a Monte Carlo study", Journal of Econometrics 96:357–393.

Bauwens, L., and P. Giot (1998), "The logarithmic ACD model: an application to the bid/ask quote process of two NYSE stocks", Working paper (CORE, Louvain-la-Neuve).

Blanchard, O.J., and P. Diamond (1994), "Ranking, unemployment duration, and wages", Review of Economic Studies 61:417–434.

Boizot, C., J.M. Robin and M. Visser (1997), "The demand for food products: An analysis of interpurchase times and purchased quantities", Working paper (CREST, Paris).

Bonnal, L., D. Fougère and A. Sérandon (1997), "Evaluating the impact of French employment policies on individual labour market histories", Review of Economic Studies 64:683–713.

Bontemps, C., J.M. Robin and G.J. Van den Berg (1999), "An empirical equilibrium job search model with search on the job and heterogeneous workers and firms", International Economic Review 40:1039–1074.

Bontemps, C., J.M. Robin and G.J. Van den Berg (2000), "Equilibrium search with continuous productivity dispersion: theory and non-parametric estimation", International Economic Review 41:305–358.

Bowlus, A.J., N.M. Kiefer and G.R. Neumann (2001), "Equilibrium search models and the transition from school to work", International Economic Review 42, forthcoming.

Bretagnolle, J., and C. Huber-Carol (1988), "Effect of omitting covariates in Cox's model for survival data", Scandinavian Journal of Statistics 15:125–138.

Burgess, S. (1989), "The estimation of structural models of unemployment duration with on-the-job search", Working paper (University of Bristol, Bristol).

Butler, J.S., K.H. Anderson and R.V. Burkhauser (1986), "Testing the relationship between work and health", Economics Letters 20:383–386.

Butler, J.S., K.H. Anderson and R.V. Burkhauser (1989), "Work and health after retirement", Review of Economics and Statistics 71:46–53.

Card, D., and D. Sullivan (1988), "Measuring the effect of subsidized training programs on movements in and out of employment", Econometrica 56:497–530.

Carling, K., and T. Jacobson (1995), "Modeling unemployment duration in a dependent competing risks framework: identification and estimation", Lifetime Data Analysis 1.

Chamberlain, G. (1985), "Heterogeneity, omitted variable bias, and duration dependence", in: J.J. Heckman and B. Singer, eds., Longitudinal Analysis of Labor Market Data (Cambridge University Press, Cambridge).

Chesher, A., and T. Lancaster (1983), "The estimation of models of labour market behaviour", Review of Economic Studies 50:609–624.

Coleman, T.S. (1990), "Unemployment behaviour: evidence from the CPS work experience survey", in: Y. Weiss and G. Fishelson, eds., Advances in the Theory and Measurement of Unemployment (Macmillan, Basingstoke).

Cox, D.R. (1972), "Regression models and life tables (with discussion)", Journal of the Royal Statistical Society Series B 34:187–220.

Dabrowska, D.M. (1987), "Non-parametric regression with censored survival time data", Scandinavian Journal of Statistics 14:181–197.

Devine, T.J., and N.M. Kiefer (1991), Empirical Labor Economics (Oxford University Press, Oxford).

Diebold, F.X., and G.D. Rudebusch (1990), "A nonparametric investigation of duration dependence in the American business cycle", Journal of Political Economy 98:596–616.

Eberwein, C., J.C. Ham and R.J. LaLonde (1997), "The impact of being offered and receiving classroom training on the employment histories of disadvantaged women: evidence from experimental data", Review of Economic Studies 64:655–682.

Elbers, C., and G. Ridder (1982), "True and spurious duration dependence: The identifiability of the proportional hazard model", Review of Economic Studies 49:403–410.

Engberg, J. (1991), "The impact of unemployment benefits on job search: structural unobserved heterogeneity and spurious spikes", Working paper (Carnegie-Mellon University, Pittsburgh).

Engberg, J., P. Gottschalk and D. Wolf (1990), "A random-effects logit model of work-welfare transitions", Journal of Econometrics 43:63–75.

Engle, R., and J. Russell (1998), "Autoregressive conditional duration; a new model for irregularly spaced transaction data", Econometrica 66:1127–1162.

Feller, W. (1971), An Introduction to Probability Theory and Its Applications II (Wiley, New York).

Fleming, T.R., and D.P. Harrington (1991), Counting Processes and Survival Analysis (Wiley, New York).

Flinn, C.J. (1996), "Labor market structure and welfare: a comparison of Italy and the U.S.", Working paper (New York University, New York).

Flinn, C.J., and J.J. Heckman (1982a), "New methods for analyzing structural models of labor force dynamics", Journal of Econometrics 18:115–168.

Flinn, C.J., and J.J. Heckman (1982b), "Models for the analysis of labor force dynamics", in: R. Basmann and G. Rhodes, eds., Advances in Econometrics, Vol. 1 (JAI Press, Greenwich).

Flinn, C.J., and J.J. Heckman (1983), "Are unemployment and out of the labor force behaviorally distinct labor force states?", Journal of Labor Economics 1:28–42.

Florens, J.P., and D. Fougère (1996), "Noncausality in continuous time", Econometrica 64:1195–1212.

Freund, J.E. (1961), "A bivariate extension of the exponential distribution", Journal of the American Statistical Association 56:971–977.

Gail, M.H., S. Wieand and S. Piantadosi (1984), "Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates", Biometrika 71:431–444.

Garcia-Perez, J.I. (1998), "Non-stationary job search with firing: a structural estimation", Working paper (CEMFI, Madrid).

Genest, C., and J. MacKay (1986), "The joy of copulas: bivariate distributions with uniform marginals", The American Statistician 40:280–283.

Gönül, F., and K. Srinivasan (1993), "Consumer purchase behavior in a frequently bought product category: estimation issues and managerial insights from a hazard function model with heterogeneity", Journal of the American Statistical Association 88:1219–1227.

Gritz, R.M. (1993), "The impact of training on the frequency and duration of employment", Journal of Econometrics 57:21–51.

Guo, G., and G. Rodríguez (1992), "Estimating a multivariate proportional hazard model for clustered data using the EM algorithm, with an application to child survival in Guatemala", Journal of the American Statistical Association 87:969–976.

Hahn, J. (1994), "The efficiency bound of the mixed proportional hazard model", Review of Economic Studies 61:607–629.

Ham, J.C., and R.J. LaLonde (1996), "The effect of sample selection and initial conditions in duration models: Evidence from experimental data on training", Econometrica 64:175–205.

Ham, J.C., and S.A. Rea (1987), "Unemployment insurance and male unemployment duration in Canada", Journal of Labor Economics 5:325–353.

Heckman, J.J. (1979), "Sample selection bias as a specification error", Econometrica 47:153–161.

Heckman, J.J. (1991), "Identifying the hand of the past: distinguishing state dependence from heterogeneity", American Economic Review 81(supplement):71–79.

Heckman, J.J., and G.J. Borjas (1980), "Does unemployment cause future unemployment? Definitions, questions and answers from a continuous time model of heterogeneity and state dependence", Economica 47:247–283.

Heckman, J.J., and B.E. Honoré (1989), "The identifiability of the competing risks model", Biometrika 76:325–330.

Heckman, J.J., and B. Singer (1984a), “Econometric duration analysis”, Journal of Econometrics 24:63–132.

Heckman, J.J., and B. Singer (1984b), “The identifiability of the proportional hazard model”, Review of Economic Studies 51:231–241.

Heckman, J.J., and B. Singer (1984c), “A method for minimizing the impact of distributional assumptions in econometric models for duration data”, Econometrica 52:271–320.

Heckman, J.J., and C.R. Taber (1994), “Econometric mixture models and more general models for unobservables in duration analysis”, Statistical Methods in Medical Research 3:279–302.

Heckman, J.J., and J.R. Walker (1987), “Using goodness of fit and other criteria to choose among competing duration models: a case study of the Hutterite data”, in: C. Clogg, ed., Sociological Methodology 1987 (American Sociological Association, Washington).

Heckman, J.J., and J.R. Walker (1990), “The relationship between wages and income and the timing and spacing of births”, Econometrica 58:1411–1441.

Heckman, J.J., V.J. Hotz and J.R. Walker (1985), “New evidence on the timing and spacing of births”, American Economic Review 75:179–184.

Heckman, J.J., R.J. LaLonde and J.A. Smith (1999), “The economics and econometrics of active labor market programs”, in: O. Ashenfelter and D. Card, eds., Handbook of Labor Economics, Vol. III (North-Holland, Amsterdam).

Honoré, B.E. (1991), “Identification results for duration models with multiple spells or time-varying covariates”, Working paper (Northwestern University, Evanston).

Honoré, B.E. (1993), “Identification results for duration models with multiple spells”, Review of Economic Studies 60:241–246.

Horowitz, J.L. (1996), “Semiparametric estimation of a regression model with an unknown transformation of the dependent variable”, Econometrica 64:103–137.

Horowitz, J.L. (1999), “Semiparametric estimation of a proportional hazard model with unobserved heterogeneity”, Econometrica 67:1001–1028.

Horvath, W.J. (1968), “A statistical model for the duration of wars and strikes”, Behavioral Science 13:24.

Hougaard, P. (1986), “A class of multivariate failure time distributions”, Biometrika 73:671–678.

Hougaard, P. (1987), “Modelling multivariate survival”, Scandinavian Journal of Statistics 14:291–304.

Hougaard, P. (1991), “Modelling heterogeneity in survival data”, Journal of Applied Probability 28: 695–701.

Hougaard, P., B. Harvald and N.V. Holm (1992a), “Measuring the similarities between the lifetimes of adult Danish twins born between 1881–1930”, Journal of the American Statistical Association 87:17–24.

Hougaard, P., B. Harvald and N.V. Holm (1992b), “Assessment of dependence in the life times of twins”, in: J.P. Klein and P.K. Goel, eds., Survival Analysis: State of the Art (Kluwer Academic Publishers, Dordrecht).

Hougaard, P., P. Myglegaard and K. Borch-Johnsen (1994), “Heterogeneity models of disease susceptibility, with an application to diabetic nephropathy”, Biometrics 50:1178–1188.

Jovanovic, B. (1984), “Wages and turnover: a parametrization of the job-matching model”, in: G.R. Neumann and N. Westergård-Nielsen, eds., Studies in Labor Market Dynamics (Springer, Heidelberg).

Kalbfleisch, J.D., and R.L. Prentice (1980), The Statistical Analysis of Failure Time Data (Wiley, New York).

Keiding, N. (1998), “Selection effects and nonproportional hazards in survival models and models for repeated events”, Working paper (University of Copenhagen, Copenhagen).

Keiding, N., P.K. Andersen and J.P. Klein (1997), “The role of frailty models and accelerated failure time models in describing heterogeneity due to omitted covariates”, Statistics in Medicine 16:215–224.

Kendall, M.G. (1962), Rank Correlation Methods (Griffin, London).

Kennan, J.F. (1985), “The duration of contract strikes in U.S. manufacturing”, Journal of Econometrics 28:5–28.

Kiefer, N.M. (1988), “Economic duration data and hazard functions”, Journal of Economic Literature 26:646–679.

Klaassen, C.A.J., and A.J. Lenstra (1998), “The information for the treatment effect in the mixed proportional hazards model vanishes”, Working paper (University of Amsterdam, Amsterdam).

Koning, P., G.J. Van den Berg, G. Ridder and K. Albæk (2000), “The relation between wages and labor market frictions: An empirical analysis based on matched worker–firm data”, in: H. Bunzel et al., eds., Panel Data and Structural Labour Market Models (North-Holland, Amsterdam).

Kortram, R.A., A.J. Lenstra, G. Ridder and A.C.M. Van Rooij (1995), “Constructive identification of the mixed proportional hazards model”, Statistica Neerlandica 49:269–281.

Lancaster, T. (1979), “Econometric methods for the duration of unemployment”, Econometrica 47: 939–956.

Lancaster, T. (1983), “Generalized residuals and heterogeneous duration models: the exponential case”, Bulletin of Economic Research 35:71–86.

Lancaster, T. (1985a), “Simultaneous equations models in applied search theory”, Journal of Econometrics 28:155–169.

Lancaster, T. (1985b), “Generalised residuals and heterogeneous duration models – with applications to the Weibull model”, Journal of Econometrics 28:113–126.

Lancaster, T. (1990), The Econometric Analysis of Transition Data (Cambridge University Press, Cambridge).

Lancaster, T. (2000a), “The incidental parameter problem since 1948”, Journal of Econometrics 95: 391–413.

Lancaster, T. (2000b), “Some econometrics of scarring”, in: Nonlinear Statistical Inference: Essays in Honor of T. Amemiya (Cambridge University Press, Cambridge).

Lenstra, A.J., and A.C.M. Van Rooij (1998), “Nonparametric estimation of the mixed proportional hazards model”, Working paper (Free University, Amsterdam).

Li, Y., J.P. Klein and M.L. Moeschberger (1993), “Effects of model misspecification in estimating covariate effects in survival analysis for small sample sizes”, Working paper (Medical College of Wisconsin, Milwaukee).

Lillard, L.A. (1993), “Simultaneous equations for hazards”, Journal of Econometrics 56:189–217.

Lillard, L.A., and C.W.A. Panis (1996), “Marital status and mortality: the role of health”, Demography 33:313–327.

Lillard, L.A., and C.W.A. Panis (1998), “Panel attrition from the PSID”, Journal of Human Resources 33:437–457.

Lindeboom, M., and M. Kerkhofs (2000), “Time patterns of work and sickness absence: unobserved workplace effects in a multi-state duration model”, Review of Economics and Statistics 82:668–684.

Lindeboom, M., and G.J. Van den Berg (1994), “Heterogeneity in bivariate duration models: the importance of the mixing distribution”, Journal of the Royal Statistical Society Series B 56:49–60.

Lindstrom, D.P. (1996), “Economic opportunity in Mexico and return migration from the United States”, Demography 33:357–374.

McCall, B.P. (1996), “The identifiability of the mixed proportional hazards model with time-varying coefficients”, Econometric Theory 12:733–738.

Meghir, C., and E. Whitehouse (1997), “Labour market transitions and retirement of men in the UK”, Journal of Econometrics 79:327–354.

Melino, A., and G.T. Sueyoshi (1990), “A simple approach to the identifiability of the proportional hazards model”, Economics Letters 33:63–68.

Meyer, B.D. (1995), “Semiparametric estimation of hazard models”, Working paper (Northwestern University, Evanston).

Mortensen, D.T. (1986), “Job search and labor market analysis”, in: O. Ashenfelter and R. Layard, eds., Handbook of Labor Economics (North-Holland, Amsterdam).

Mortensen, D.T., and C.A. Pissarides (1999), "New developments in models of search in the labor market", in O. Ashenfelter and D. Card, eds., Handbook of Labor Economics, Vol. III (North-Holland, Amsterdam).

Moscarini, G. (1997), "Unobserved heterogeneity and unemployment duration: a fallacy of composition", Working paper (Yale University, New Haven).

Mroz, T., and D. Weir (1990), "Structural change in life cycle fertility during the fertility transition: France before and after the revolution of 1789", Population Studies 44:61–87.

Narendranathan, W. (1993), "Job search in a dynamic environment – an empirical analysis", Oxford Economic Papers 45:1–22.

Narendranathan, W., and S.J. Nickell (1985), "Modelling the process of job search", Journal of Econometrics 28:28–49.

Neumann, G.R. (1997), "Search models and duration data", in: M.H. Pesaran, ed., Handbook of Applied Econometrics: Microeconometrics (Basil Blackwell, Oxford).

Newman, J.L., and C.E. McCullogh (1984), "A hazard rate approach to the timing of births", Econometrica 52:939–961.

Ng, E.T.M., and R.J. Cook (1997), "Modeling two-state disease processes with random effects", Lifetime Data Analysis 3:315–335.

Nickell, S.J. (1979), "Estimating the probability of leaving unemployment", Econometrica 47:1249–1266.

Nielsen, G.G., R.D. Gill, P.K. Andersen and T.I.A. Sørensen (1992), "A counting process approach to maximum likelihood estimation in frailty models", Scandinavian Journal of Statistics 19:25–43.

Oakes, D. (1989), "Bivariate survival models induced by frailties", Journal of the American Statistical Association 84:487–493.

Omori, Y. (1997), "Stigma effects of nonemployment", Economic Inquiry 35:394–416.

Pakes, A., and M. Schankerman (1984), "The rate of obsolescence of knowledge, research gestation lags, and the private rate of return to research resources", in: Z. Griliches, ed., Patents, R&D and Productivity (University of Chicago Press, Chicago).

Pelz, C.J., and J.P. Klein (1996), "Analysis of survival data: a comparison of three major statistical packages (SAS, SPSS, BMDP)", Working paper (Medical College of Wisconsin, Milwaukee).

Petersen, J.H. (1996), "A litter frailty model", Working paper (University of Copenhagen, Copenhagen).

Pissarides, C.A. (1990), Equilibrium Unemployment Theory (Basil Blackwell, Oxford).

Ridder, G. (1984), "The distribution of single-spell duration data", in: G.R. Neumann and N. Westergård-Nielsen, eds., Studies in Labor Market Dynamics (Springer, Heidelberg).

Ridder, G. (1987), "The sensitivity of duration models to misspecified unobserved heterogeneity and duration dependence", Working paper (Groningen University, Groningen).

Ridder, G. (1990), "The non-parametric identification of generalized accelerated failure-time models", Review of Economic Studies 57:167–182.

Ridder, G., and I. Tunalı (1999), "Stratified partial likelihood estimation", Journal of Econometrics 92:193–232.

Ridder, G., and G.J. Van den Berg (1997), "Empirical equilibrium search models", in: D.M. Kreps and K.F. Wallis, eds., Advances in Economics and Econometrics: Theory and Applications (Cambridge University Press, Cambridge).

Ridder, G., and G.J. Van den Berg (1998), "Estimating measures of labor market imperfection for five OECD countries, using aggregate data in an equilibrium search framework", Working paper (Free University, Amsterdam).

Robins, J.M. (1998), "Structural nested failure time models", in: P. Armitage and T. Colton, eds., The Encyclopedia of Biostatistics (Wiley, Chichester).

Robins, J.M., and S. Greenland (1989), "The probability of causation under a stochastic model for individual risk", Biometrics 45:1125–1138.

Rosholm, M. (1997), "The risk of marginalization in the labour market: application of a three state dependent competing risks duration model", Working paper (University of Aarhus, Aarhus).

Rust, J. (1994), "Structural estimation of Markov decision processes", in: R.F. Engle and D.L. McFadden, eds., Handbook of Econometrics, Vol. IV (North-Holland, Amsterdam).

Ryu, K. (1993), "Structural duration analysis of management data", Journal of Econometrics 57:91–115.

Sastry, N. (1997), "A nested frailty model for survival data, with an application to the study of child survival in Northeast Brazil", Journal of the American Statistical Association 92:426–435.

Snyder, D.L., and M.I. Miller (1991), Random Point Processes in Time and Space (Springer, Heidelberg).

Solomon, P.J. (1984), "Effect of misspecification of regression models in the analysis of survival data", Biometrika 71:291–298; correction: 1986, 73:245.

Solon, G. (1985), "Work incentive effects of taxing unemployment benefits", Econometrica 53:295–306.

Thomas, J.M. (1998), "The role of selective job search in UK unemployment", Economic Journal 108:646–664.

Trussell, J., and T. Richards (1985), "Correcting for unmeasured heterogeneity in hazard models using the Heckman–Singer procedure", in: N. Tuma, ed., Sociological Methodology 1985 (Jossey-Bass, San Francisco).

Van den Berg, G.J. (1990a), "Nonstationarity in job search theory", Review of Economic Studies 57:255–277.

Van den Berg, G.J. (1990b), "Search behaviour, transitions to nonparticipation and the duration of unemployment", Economic Journal 100:842–865.

Van den Berg, G.J. (1990c), "The effect of an increase of the rate of arrival of job offers on the duration of unemployment", Working paper (Groningen University, Groningen).

Van den Berg, G.J. (1992), "Nonparametric tests for unobserved heterogeneity in duration models", Working paper (Free University, Amsterdam).

Van den Berg, G.J. (1995), "Explicit expressions for the reservation wage path and the unemployment duration density in nonstationary job search models", Labour Economics 2:187–198.

Van den Berg, G.J. (1997), "Association measures for durations in bivariate hazard rate models", Journal of Econometrics 79:221–245.

Van den Berg, G.J. (1999), "Empirical inference with equilibrium search models of the labor market", Economic Journal 109:F283–F306.

Van den Berg, G.J., and M. Lindeboom (1998), "Attrition in panel survey data and the estimation of multi-state labor market models", Journal of Human Resources 33:458–478.

Van den Berg, G.J., and G. Ridder (1998), "An empirical equilibrium search model of the labor market", Econometrica 66:1183–1221.

Van den Berg, G.J., and J.C. Van Ours (1996), "Unemployment dynamics and duration dependence", Journal of Labor Economics 14:100–125.

Van den Berg, G.J., M. Lindeboom and G. Ridder (1994), "Attrition in longitudinal panel data, and the empirical analysis of dynamic labour market behaviour", Journal of Applied Econometrics 9:421–435.

Van den Berg, G.J., B. Van der Klaauw and J.C. Van Ours (1998), "Punitive sanctions and the transition rate from welfare to work", Working paper (Tinbergen Institute, Amsterdam).

Van den Berg, G.J., A. Holm and J.C. Van Ours (2001), "Do stepping-stone jobs exist? Early career paths in the medical profession", Journal of Population Economics, forthcoming.

Van der Klaauw, W. (1996), "Female labour supply and marital status decisions: a life-cycle model", Review of Economic Studies 63:199–235.

Vaupel, J.W., K.G. Manton and E. Stallard (1979), "The impact of heterogeneity in individual frailty on the dynamics of mortality", Demography 16:439–454.

Vilcassim, N.J., and D.C. Jain (1991), "Modeling purchase-timing and brand-switching behavior incorporating explanatory variables and unobserved heterogeneity", Journal of Marketing Research 28:29–41.

Visser, M. (1996), "Nonparametric estimation of the bivariate survival function with an application to vertically transmitted AIDS", Biometrika 83:507–518.

Wang, S.T., J.P. Klein and M.L. Moeschberger (1995), "Semi-parametric estimation of covariate effects using the Positive Stable frailty model", Applied Stochastic Models and Data Analysis 11:121–133.

Wolpin, K.I. (1987), "Estimating a structural job search model: the transition from school to work", Econometrica 55:801–818.

Wolpin, K.I. (1995), Empirical methods for the study of labor force dynamics (Harwood Academic Publishers, Luxembourg).

Woutersen, T.M. (2000), "Consistent estimators for panel duration data with endogenous censoring and endogenous regressors", Working paper (Brown University, Providence).

Xue, X., and R. Brookmeyer (1996), "Bivariate frailty model for the analysis of multivariate survival time", Lifetime Data Analysis 2:277–289.

Yamaguchi, K. (1986), "Alternative approaches to unobserved heterogeneity in the analysis of repeatable events", in: N.B. Tuma, ed., Sociological Methodology 1986 (Jossey-Bass, Washington, DC).

Yamaguchi, K. (1991), Event History Analysis (Sage, Newbury Park).

Yashin, A.I., and I.A. Iachine (1997), "How frailty models can be used in evaluating longevity limits", Demography 34:31–48.

Yashin, A.I., and I.A. Iachine (1999), "What difference does the dependence between durations make? Insights for population studies of aging", Lifetime Data Analysis 5:5–22.

Yoon, B.J. (1981), "A model of unemployment duration with variable search intensity", Review of Economics and Statistics 63:599–609.

Yoon, B.J. (1985), "A non-stationary hazard function of leaving unemployment for employment", Economics Letters 17:171–175.

Zahl, P.H. (1997), "Frailty modelling for the excess hazard", Statistics in Medicine 16:1573–1585.

Part 12

COMPUTATIONAL METHODS IN ECONOMETRICS

This Page Intentionally Left Blank

Chapter 56

COMPUTATIONALLY INTENSIVE METHODS FOR INTEGRATION IN ECONOMETRICS*

JOHN GEWEKE

University of Iowa

MICHAEL KEANE

New York University

Contents

Abstract 3465
Keywords 3465
1. Introduction 3466
2. Monte Carlo methods of integral approximation 3468
2.1. Independence sampling 3470
2.2. The Gibbs sampler 3472
2.3. The Hastings–Metropolis algorithm 3474
2.4. Some Markov chain Monte Carlo theory 3476
2.5. Metropolis within Gibbs 3479
2.6. Assessing numerical accuracy in Markov chain Monte Carlo 3479
3. Approximate solution of discrete dynamic optimization problems 3481
4. Classical simulation estimation of the multinomial probit model 3494
5. Univariate latent linear models 3506
5.1. An overview of the univariate latent linear model 3507
5.1.1. Distribution of disturbances 3507
5.1.2. Observable outcomes 3508
5.1.3. Prior distributions 3508
5.1.4. Existence of the posterior 3509
5.1.5. MCMC algorithm for inference 3510
5.1.6. Marginal likelihoods 3511
5.2. Some evidence from artificial data 3511
5.2.1. Parameter posterior moments 3512
5.2.2. Marginal likelihood approximations 3516

* The authors gratefully acknowledge financial support from National Science Foundation grants SBR-9511280, SBR-9731037, SES-9814342, and SES-9819444.

Handbook of Econometrics, Volume 5, Edited by J.J. Heckman and E. Leamer

5.3. Some evidence from real data 3518

6. Multivariate latent linear models 3518

6.1. An overview of the multivariate latent linear model 3519

6.1.1. Linear restrictions in the multivariate latent linear model 3519

6.1.2. Distribution of disturbances 3520

6.1.3. Observed outcomes 3521

6.1.4. Prior distributions 3522

6.1.5. Existence of the posterior distribution 3524

6.1.6. MCMC algorithm for inference 3524

6.1.7. Marginal likelihoods 3525

6.2. Some evidence from artificial data 3525

6.2.1. Multivariate linear model 3525

6.2.2. Degenerate multivariate linear model 3528

6.2.3. Multiple discrete choice model 3528

6.2.4. Continuous selection model 3530

6.2.5. Discrete choice selection model 3533

6.2.6. Marginal likelihood approximations 3535

6.2.7. Prospects for applications and future development 3537

7. Bayesian inference for a dynamic discrete choice model 3538

Appendix A. The full univariate latent linear model 3548

Appendix B. The full multivariate latent linear model 3555

References 3564

Abstract

Until recently, inference in many interesting models was precluded by the requirement of high dimensional integration. But dramatic increases in computer speed, and the recent development of new algorithms that permit accurate Monte Carlo evaluation of high dimensional integrals, have greatly expanded the range of models that can be considered. This chapter presents the methodology for several of the most important Monte Carlo methods, supplemented by a set of concrete examples that show how the methods are used.

Some of the examples are new to the econometrics literature. They include inference in multinomial discrete choice models and selection models in which the standard normality assumption is relaxed in favor of a multivariate mixture of normals assumption. Several Monte Carlo experiments indicate that these methods are successful at identifying departures from normality when they are present. Throughout the chapter the focus is on inference in parametric models that permit rich variation in the distribution of disturbances.

The chapter first discusses Monte Carlo methods for the evaluation of high dimensional integrals, including integral simulators like the GHK method, and Markov Chain Monte Carlo methods like Gibbs sampling and the Metropolis–Hastings algorithm. It then turns to methods for approximating solutions to discrete choice dynamic optimization problems, including the methods developed by Keane and Wolpin, and Rust, as well as methods for circumventing the integration problem entirely, such as the approach of Geweke and Keane. The rest of the chapter deals with specific examples: classical simulation estimation for multinomial probit models, both in the cross sectional and panel data contexts; univariate and multivariate latent linear models; and Bayesian inference in dynamic discrete choice models in which the future component of the value function is replaced by a flexible polynomial.

Keywords

Bayesian inference, discrete choice, dynamic optimization, integration, Markov chain Monte Carlo, multinomial probit, normal mixtures, selection models

JEL classification: Primary C15, Secondary C11

1. Introduction

There are many inferential problems in econometrics for which the evaluation of high dimensional integrals is essential. The example most familiar to econometricians is perhaps the classical estimation of discrete choice models, such as multinomial probit (MNP), when the number of alternatives is large. Construction of the likelihood function for a discrete choice model requires evaluation of the choice probabilities generated by the model. Those choice probabilities take the form of integrals over regions of the disturbance space such that particular choices are generated. In familiar models like MNP, if there are J alternatives in the choice set the disturbance space is $J-1$ dimensional, and evaluation of $J-1$ dimensional integrals is in general required. For small J this is computationally feasible using highly accurate series expansions or quadrature methods, which we will not discuss in this chapter. But for large J such highly accurate numerical methods are computationally infeasible, and the Monte Carlo methods which are the subject of this chapter become essential[1].

While discrete choice models have received the most attention to date, there are many other and no less important instances in econometrics where the need for high dimensional integration arises. A prime example is in Bayesian inference where evaluation of posterior distributions of model parameters (as well as other posterior moments of interest) often requires high dimensional integration even when evaluation of the likelihood does not. For instance, consider models with latent variables or other nuisance parameters that must be integrated out to form the marginal posterior of the parameters of interest. The order of integration required to form this marginal posterior will equal the number of latent variables and/or nuisance parameters. Another example is the evaluation of marginal likelihoods – the integral of the likelihood with respect to the prior density of the model parameters – which are critical in Bayesian analysis for comparing the plausibility of different models. Clearly these integrals have dimension equal to the number of model parameters, and, except in very special cases where they have closed forms, they must be evaluated numerically.

Another important area in which difficult integration problems arise is in models of economic agents who solve optimization problems that include discrete control variables. Inference in such models generally requires that the econometrician solve, at many points in the feasible parameter space, the optimization problem assumed to be solved by the agents. This requires that the value functions at each point in the state space of the problem be calculated, and these value functions are typically high dimensional integrals.

Yet another area of interest is in state space models, leading examples of which are stochastic volatility models. In such models the variance of the stochastic terms is a

[1] The order of integration that is feasible using quadrature methods is a function of the speed of available computers. At the present time the boundary appears to be at about $J = 3$ or 4, although there is some controversy on this point.

latent variable which itself follows a stochastic process. If this process exhibits serial correlation, the likelihood for any history of the dependent variables takes the form of a T dimensional integral over the density of the time period specific values of the variance term, where T is the number of time periods.

Until recently, inference in many interesting models was precluded by the requirement of high dimensional integration. But dramatic increases in computer speed, and the recent development of new algorithms that permit accurate Monte Carlo evaluation of high dimensional integrals, have greatly expanded the range of models that can be considered. This chapter will present the methodology for several of the most important Monte Carlo methods, supplemented by a set of concrete examples that show how the methods are used.

Some of the examples we present are new to the econometrics literature and in our view rather significant. For instance, we show how to conduct inference in multinomial discrete choice models and Heckman selection models in which the standard normality assumption is relaxed in favor of a multivariate mixture of normals assumption. Several Monte Carlo experiments indicate that these methods are successful in identifying departures from normality when they are present.

Throughout this chapter our focus is on inference in parametric models that permit rich variation in the distribution of disturbances. We take this approach for several reasons. First, there are many instances in economics in which a complete model – including distributions for all the relevant stochastic terms – is needed to predict behavior. One example is a risk averse, expected utility maximizing agent choosing between two risky income streams: expected utility will depend on the whole distribution of shocks to each income stream. Another is discrete choice: the effect on choice probabilities if a covariate or policy variable is changed will depend on the entire probability distribution of the shock to latent utility. A second reason for taking this approach is that it is natural to use Bayesian methods in connection with these likelihood based procedures, and to use Bayes factors to discriminate between different models of interest and to provide a practical but well grounded guide for the number of parameters to be used in a given application. These procedures lead to high dimensional integration problems. The good news is that methods developed in recent years to attack these problems have led to practical strategies within the reach of most applied econometricians.

The outline of this chapter is as follows. In Section 2 we discuss Monte Carlo methods for the evaluation of high dimensional integrals, including integral simulators like the GHK method, and Markov Chain Monte Carlo methods like Gibbs sampling and the Metropolis–Hastings algorithm. Section 3 discusses methods for approximating solutions to discrete choice dynamic optimization problems, including the methods developed by Keane and Wolpin (1994) and Rust (1997), as well as methods for circumventing the integration problem entirely, such as Geweke and Keane (1995). Sections 4 through 7 discuss the application of the integration methods from Sections 2 and 3 to a number of inferential problems. Section 4 deals with classical simulation estimation for multinomial probit models, both in the cross sectional and panel data contexts. Section 5 discusses Bayesian inference in univariate latent linear models. The emphasis is on computational methods that allow the normality assumption to be relaxed in favor of more flexible mixture of normals and Student-t error structures. Section 6 extends the methods of Section 5 to a multi-equation setting. Section 7 considers Bayesian inference in a dynamic discrete choice model in which the future component of the value function is replaced by a flexible polynomial.

2. Monte Carlo methods of integral approximation

A generic problem that arises often in econometrics is to evaluate an integral of the form

$$E\,[g(\boldsymbol{x})] = \int_S g(\boldsymbol{x})p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}, \tag{2.1}$$

where $\boldsymbol{x} \in S \subseteq \Re^k$ and $p(\cdot)$ is a probability density function (p.d.f.) with support on S. Two leading circumstances in which these problems arise are latent or limited dependent variable models (regardless of the method of inference) and Bayesian inference (regardless of the kind of model).

The multinomial probit model provides a classic example of this problem. Suppose that an individual derives utility u_j from choices $j = 1, \ldots, J-1$, utility 0 from choice J, and $\boldsymbol{u} = (u_1, \ldots, u_{J-1})' \sim N(\mu, \Sigma)$. Then the probability of making choice j is of the form (2.1), with $k = J-1$, $\boldsymbol{x} = \boldsymbol{u}$, $g(\boldsymbol{u}) = 1$ if and only if $u_j \geqslant u_i (i = 1, \ldots, J-1)$ and $u_J \geqslant 0$, and $p(\boldsymbol{u})$ is the p.d.f. of the $N(\mu, \Sigma)$ density. Perhaps the most obvious Monte Carlo method of integral approximation in this example is to draw $\boldsymbol{u}^{(m)} \overset{\text{i.i.d.}}{\sim} N(\mu, \Sigma)$ $(m = 1, \ldots, M)$ and set $\bar{g}^{(M)} = M^{-1}\sum_{m=1}^{M} g(\boldsymbol{u}^{(m)})$. Then $\bar{g}^{(M)} \xrightarrow{a.s.} \bar{g} \equiv \mathrm{E}[g(\boldsymbol{x})]$ and $M^{1/2}\left(\bar{g}^{(M)} - \bar{g}\right) \xrightarrow{d} N[0, \bar{g}(1-\bar{g})]$. But this *crude frequency simulator,* discussed by Albright, Lerman and Manski (1977) and Lerman and Manski (1981), has well known practical problems: for example, if a nonlinear function of the choice probability is important, as it is in evaluating a likelihood function, then a great many simulations may be required, and if choice probability is small then the approximation of the log choice probability may be quite difficult. We shall revisit this problem more than once in this chapter.

Bayesian inference begins from a data density $p(\boldsymbol{y}|\theta_A, A)$ in which $\boldsymbol{y} \in \Re^T$ denotes the observable data, $\theta_A \in \Theta_A \subseteq \Re^k$ is a vector of unknown parameters, and A indexes the model. Given a prior density $p(\theta_A|A)$, the model implies the joint density $p(\boldsymbol{y}, \theta_A|A) = p(\theta_A|A)p(\boldsymbol{y}|\theta_A, A)$. Then the *marginal likelihood* of the observed data $\boldsymbol{y} = \boldsymbol{y}^o$ conditional on A is

$$p(\boldsymbol{y}^o|A) = \int_{\Theta_A} p(\boldsymbol{y}^o|\theta_A, A)p(\theta_A, A)\,\mathrm{d}\nu(\theta_A), \tag{2.2}$$

and the *posterior density* of θ_A given the data $\boldsymbol{y}^o$ in model A is

$$p(\theta_A|\boldsymbol{y}^o, A) = p(\boldsymbol{y}^o|\theta_A, A)p(\theta_A|A)/p(\boldsymbol{y}^o|A). \tag{2.3}$$

Typically there is a vector of interest, separate from the particular model A, of the form $\omega \in \Omega \subseteq \Re^{\ell}$. For example, ω could be a future event, the consequence of a conjectured change in policy, an aspect of tastes like risk aversion, or an aspect of technology like returns to scale. If the model A has implications for this vector, these implications can be expressed $p(\omega|\boldsymbol{y}, \theta_A, A)$. The expectation of any function $h(\omega)$ given the data $\boldsymbol{y}^o$, conditional on the model A is

$$\mathrm{E}[h(\omega)|\boldsymbol{y}^o, A] = \int_{\Omega}\int_{\Theta_A} h(\omega)p(\omega|\boldsymbol{y}^o, \theta_A, A)p(\theta_A|\boldsymbol{y}^o, A). \tag{2.4}$$

Expression (2.4), which includes most Bayesian inference, implicitly carries forward the integration problem in Equation 2.2, by means of Equation 2.3, as well as the integration shown in Equation 2.4.

The most obvious method of Monte Carlo integration in Equation 2.4 would be to draw $\theta_A^{(m)} \sim \mathrm{p}(\theta_A|A)$, $\boldsymbol{y}^{(m)} \sim p(\boldsymbol{y}|\theta_A^{(m)}, A)$, $\omega^{(m)} \sim p(\omega|\boldsymbol{y}^{(m)}, \theta_A^{(m)}, A)$. If $\boldsymbol{y}$ is discrete and $p(\boldsymbol{y} = \boldsymbol{y}^o|A)$ is not too small, this procedure is practical. But if this probability is small, or if $\boldsymbol{y}$ is continuous so that with probability one $\boldsymbol{y}^{(m)} \neq \boldsymbol{y}^o$ for all m, then Equation 2.4 cannot be approximated in this way. Kernel density methods can, in principle, cope with this problem. Let $K(u)$ be a function with property $\int_{-\infty}^{\infty} K(u)\,\mathrm{d}u = 1$ and let $d(\boldsymbol{y}_1,\boldsymbol{y}_2)$ be a measure of distance between any two $\boldsymbol{y}_i \in \Re^T$. The *kernel-smoothed frequency simulator* approximates Equation (2.4) by $cM^{\alpha}\sum_{m=1}^{M} K[d(\boldsymbol{y}^o,\boldsymbol{y}^{(m)})/cM^{\alpha}]h(\omega^{(m)})$ where $0 < \alpha < \frac{1}{2}$, $\alpha = \frac{1}{5}$ gives optimal results in some circumstances, and the choice of c is problematic. The approximation is consistent in simulation size M. [Tapia and Thompson (1978, Chapter 2) provides analytic detail and practical guidance.] The real difficulty with this procedure is that kernel density methods are only practical up to dimension 3 or 4; beyond that, the number of simulations M required is too great. Of course, T is typically much larger than 3 or 4.

An approach that eliminates the dimensionality of $\boldsymbol{y}$ as a stumbling block is to take $\theta_A^{(m)} \sim p(\theta_A|A)$, draw $\omega^{(m)} \sim p(\omega|\boldsymbol{y}^o, \theta_A^{(m)}, A)$, and then average $h(\omega^{(m)})$ weighted by the likelihood function $p(\boldsymbol{y}^o|\theta_A, A)$. This method is an example of *importance sampling*, because it makes draws from an incorrect distribution (here, $p(\theta_A|A)$ rather than $p(\theta_A|\boldsymbol{y}^o, A)$), and makes a weighting correction (here, $p(\boldsymbol{y}^o|\theta_A, A)$) to adjust for the discrepancy. Importance sampling was first suggested in Bayesian econometrics by Kloek and van Dijk (1978), and is further treated by Geweke (1989, 1996). It is practical only in situations where θ_A has only a few elements and the likelihood function is not too concentrated relative to the prior density. In the vastly more common situation in which these conditions are not met, practically all draws $\theta_A^{(m)} \sim p(\theta_A|A)$ are far from the support of the likelihood function and nearly all the weights $p(\boldsymbol{y}^o|\theta_A^{(m)}, A)$ are negligible.

These two examples share the common feature that the obvious simulations have good large sample simulation properties, but are impractical given foreseeable computing power. Several approaches have been developed that successfully cope with these and related problems. One approach is to find independent, identically distributed simulation schemes that concentrate on the support of the distributions, thus addressing the problem directly. An example of this approach, applied to the first example, is given in Section 2.1. Often it is impossible to find such i.i.d. simulations, and in this circumstance Markov chain Monte Carlo simulates have proven very effective. The rest of this section is devoted to this approach to Monte Carlo integration.

What follows covers only selected points. More detailed recent surveys include Hajivassiliou and Ruud (1994) and Geweke (1996, 1999). Sections 2.2 through 2.7 are based closely on Geweke (1999, Section 3).

2.1. Independence sampling

In the generic problem (2.1) the crude frequency simulator is $\boldsymbol{x}^{(m)} \sim p(\boldsymbol{x})$, $g^{(m)} = g(\boldsymbol{x}^{(m)})$ $(m = 1, \ldots, M)$, and the simulation approximation of $\bar{g} = \mathrm{E}[g(\boldsymbol{x})]$ is $\bar{g}^{(M)} = M^{-1}\sum_{m=1}^{M} g^{(m)}$. The accuracy of this approximation is governed by $M^{1/2}(\bar{g}^{(m)} - \bar{g}) \xrightarrow{d} N(0, \sigma^2)$ so long as the second central moment $\sigma^2 = \mathrm{var}_p(g) = \int_S [g(\boldsymbol{x}) - \bar{g}]^2 p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$ exists. The difficulty pointed out in the two examples presented above is the size of σ^2, relative to $\bar{g}$ and the purposes at hand.

More sophisticated independence Monte Carlo approximations to $\bar{g}$ can be constructed by finding g^* and p^* such that p^* is a p.d.f. from which it is practical to draw i.i.d. synthetic variates, and $\mathrm{E}[g(\boldsymbol{x})] = \int_S g^*(\boldsymbol{x})p^*(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$ (compare Equation 2.1). The accuracy of approximation is governed by the same central limit theorem but now with $\sigma_*^2 = \mathrm{var}_{p^*}(g^*)$ in place of σ^2. The choice

$$p^*(\boldsymbol{x}) = p(\boldsymbol{x})\,g(\boldsymbol{x})/\bar{g}, \quad g^*(\boldsymbol{x}) = \bar{g},$$

would drive σ_*^2 to zero. This choice is impractical, because it requires us to solve the problem analytically in order to construct the simulator and leaves open the question of how to draw i.i.d. synthetic variates from p^*. However, it correctly suggests that very large increases in the accuracy of the Monte Carlo approximation may often be attained in this way. [For a general approach to reducing σ_*^2, see Geweke (1988)].

To see how more sophisticated independence sampling schemes can be constructed, return to the multinomial probit example introduced at the start of this section. In order to motivate the class of *iterative conditional probability simulators,* consider the general situation in which $\boldsymbol{x}' = (x_1, \ldots, x_n)$ has p.d.f. $p(x_1, \ldots, x_n)$. We wish to find $P[a_i \leqslant x_i \leqslant b_i \ (i = 1, \ldots, n)]$ for some set of pairs of extended real numbers (a_i, b_i) where $a_i < b_i$ $(i = 1, \ldots, n)$. A specific example of such a problem is the multinomial probit model above. The specific task of evaluating the probability of choice j, for $j < J$, corresponds to taking $n = J - 1$ and defining $\boldsymbol{A}$: $n \times n$ by setting $a_{ii} = -1$ $(i \neq j)$ and $a_{ij} = 1$ for $i = 1, \ldots, n$ and setting all other elements of $\boldsymbol{A}$ to 0. Then take $\boldsymbol{x} = \boldsymbol{A}\boldsymbol{u}$,

$a_i = 0$ and $b_i = +\infty$ ($i = 1, \ldots, n$), and take f to be the p.d.f. of the $N(\boldsymbol{A}\mu, \boldsymbol{A}\Sigma\boldsymbol{A}')$ distribution.

Let the random variable $\tilde{x}_1$ be drawn from the marginal distribution of x_1 subject to $a_1 \leqslant x_1 \leqslant b_1$, and let $\tilde{x}_j$ be drawn from the conditional distribution of x_j given $\tilde{x}_1, \ldots, \tilde{x}_{j-1}$ subject to $a_j \leqslant x_j \leqslant b_j$ ($j = 2, \ldots, n-1$).

Theorem 2.1.1.

$$\begin{aligned} &P[a_i \leqslant x_i \leqslant b_i \ (i = 1, \ldots, n)] \\ &\quad = P(a_1 \leqslant x_1 \leqslant b_1) \cdot \mathrm{E}\left[\prod_{i=2}^{n} P(a_i \leqslant x_i \leqslant b_i|\tilde{x}_1, \ldots, \tilde{x}_{i-1})\right]. \end{aligned}$$

Proof: Let

$$p^*(\tilde{x}_j|\tilde{x}_1, \ldots, \tilde{x}_{j-1}) = p(x_j|\tilde{x}_1, \ldots, \tilde{x}_{j-1})\chi_{(a_j, b_j)}(x_j) / \int_{a_j}^{b_j} p(x_j\,|\tilde{x}_1, \ldots, \tilde{x}_{j-1})\,\mathrm{d}x_j,$$

denote the density of $\tilde{x}_j$ given $\tilde{x}_1, \ldots, \tilde{x}_{j-1}$ subject to $a_j \leqslant x_j \leqslant b_j$.

$$\begin{aligned} &E\left[\prod_{i=2}^{n} P(a_i \leqslant x_i \leqslant b_i|\tilde{x}_1, \ldots, \tilde{x}_{i-1})\right] \\ &\quad = \int_{a_1}^{b_1} \cdots \int_{a_{n-1}}^{b_{n-1}} P(a_2 \leqslant x_2 \leqslant b_2|\tilde{x}_1) P(a_3 \leqslant x_3 \leqslant b_3|\tilde{x}_1, \tilde{x}_2) \qquad (2.5) \\ &\cdots P(a_{n-1} \leqslant x_{n-1} \leqslant b_{n-1}|\tilde{x}_1, \ldots, \tilde{x}_{n-2}) P(a_n \leqslant x_n \leqslant b_n|\tilde{x}_1, \ldots, \tilde{x}_{n-1}) \\ &\cdot p^*(\tilde{x}_1) p^*(\tilde{x}_2|\tilde{x}_1) \cdots p^*(\tilde{x}_{n-2}|\tilde{x}_1, \ldots, \tilde{x}_{n-3}) p^*(\tilde{x}_{n-1}|\tilde{x}_1, \ldots, \tilde{x}_{n-2}) \\ &\mathrm{d}\tilde{x}_1 \mathrm{d}\tilde{x}_2 \cdots \mathrm{d}\tilde{x}_{n-2} \mathrm{d}\tilde{x}_{n-1}. \end{aligned}$$

The integral over $\tilde{x}_{n-1}$ in the portion of expression (2.5) involving that term is

$$\begin{aligned} &\int_{a_{n-1}}^{b_{n-1}} P(a_n \leqslant x_n \leqslant b_n|\tilde{x}_1, \ldots, \tilde{x}_{n-2}, \tilde{x}_{n-1}) p^*(\tilde{x}_{n-1}|\tilde{x}_1, \ldots, \tilde{x}_{n-2})\,\mathrm{d}\tilde{x}_{n-1} \\ &\quad = P(a_n \leqslant x_n \leqslant b_n|\tilde{x}_1, \ldots, \tilde{x}_{n-2}; a_{n-1} \leqslant x_{n-1} \leqslant b_{n-1}). \end{aligned}$$

Substituting the last expression, the integral over $\tilde{x}_{n-2}$ and $\tilde{x}_{n-1}$ in the portion of expression (2.5) involving those terms is

$$\begin{aligned} &\int_{a_{n-2}}^{b_{n-2}} P(a_n \leqslant x_n \leqslant b_n|\tilde{x}_1, \ldots, \tilde{x}_{n-2};\ a_{n-1} \leqslant x_{n-1} \leqslant b_{n-1}) \\ &\qquad \cdot P(a_{n-1} \leqslant x_{n-1} \leqslant b_n|\tilde{x}_1, \ldots, \tilde{x}_{n-2}) p^*(\tilde{x}_{n-2}|\tilde{x}_1, \ldots, \tilde{x}_{n-3})\,\mathrm{d}\tilde{x}_{n-2} \\ &\quad = P\left[a_i \leqslant x_i \leqslant b_i (i = n-1, n)|\tilde{x}_1, \ldots, \tilde{x}_{n-3};\ a_{n-2} \leqslant x_{n-2} \leqslant b_{n-2}\right]. \end{aligned}$$

Proceeding in this way, the integral over $\tilde{x}_j, \ldots, \tilde{x}_{n-1}$ in the portion of expression (1) involving those terms is

$$\begin{aligned}&\textstyle\int_{a_j}^{b_j} P\left(a_i \leqslant x_i \leqslant b_i(i = j+2, \ldots, n)|\tilde{x}_1, \ldots, \tilde{x}_j;\ a_{j+1} \leqslant x_{j+1} \leqslant b_{j+1}\right)\\ &\quad \cdot P\left(a_{j+1} \leqslant x_{j+1} \leqslant b_{j+1}|\tilde{x}_1, \ldots, \tilde{x}_j\right) p^*\left(\tilde{x}_j|\tilde{x}_1, \ldots, \tilde{x}_{j-1}\right)\, \mathrm{d}\tilde{x}_j\\ &\quad = P\left[a_i \leqslant x_i \leqslant b_i(i = j+1, \ldots, n)|\tilde{x}_1, \ldots, \tilde{x}_{j-1}; a_j \leqslant x_j \leqslant b_j\right].\end{aligned}$$

At $j = 1$ the last expression becomes $P[a_i \leqslant x_i \leqslant b_i\ (i = 2, \ldots, n)]$; multiplying by $P(a_1 \leqslant x_1 \leqslant b_1)$ gives the result. □

This result can be used to construct a practical simulator so long as it is easy to evaluate $P(a_i \leqslant x_i \leqslant b_i)$ and to draw $\tilde{x}_j$ from the conditional distribution of x_j given $\tilde{x}_1, \ldots, \tilde{x}_{j-1}$ subject to $a_j \leqslant x_j \leqslant b_j (j = 2, \ldots, n-1)$. Iteration m of the algorithm consists of drawing $\tilde{x}_1^{(m)}, \ldots, \tilde{x}_{n-1}^{(m)}$ in succession, computing $P\left(a_j \leqslant x_j \leqslant b_j|\tilde{x}_1^{(m)}, \ldots, \tilde{x}_{j-1}^{(m)}\right)$ $(j = 2, \ldots, n)$, and taking $g^{(m)} = P(a_1 \leqslant x_1 \leqslant b_1) \prod_{j=2}^{n} P\left(a_j \leqslant x_j \leqslant b_j|\tilde{x}_1^{(m)}, \ldots, \tilde{x}_{j-1}^{(m)}\right)$. The approximation of $g = P\left[a_i \leqslant x_i \leqslant b_i(i = 1, \ldots, n)\right]$ after M steps of this algorithm is $\bar{g}^{(M)} = M^{-1} \sum_{m=1}^{M} g^{(m)}$. Using the central limit theorem, the standard error of approximation for g is $M^{-1/2} \left[\bar{g}^{(M)}\left(1 - \bar{g}^{(M)}\right)\right]^{1/2}$ and that for $\log(g)$ is $M^{-1/2} \left[\left(1 - \bar{g}^{(M)}\right)/\bar{g}^{(M)}\right]^{1/2}$ If the x_i are mutually independent, then the error of approximation is zero. This limiting case establishes a class of situations in which the iterative conditional probability simulator provides dramatic increases in accuracy over the crude frequency simulator: the variates x_i are nearly independent, and the probability g is small. In both the crude frequency and iterative conditional probability simulators, $\mathrm{E}\left[g^{(m)}\right] = g$ and $0 \leqslant g^{(m)} \leqslant 1$. The crude frequency simulator *maximizes* $\mathrm{var}\left[g^{(m)}\right]$ over this class of independent $g^{(m)}$. Therefore, the iterative conditional probability simulator is always more efficient than the crude frequency simulator, in this class of problems, given the same number of iterations M. Whether it is more cost efficient may be an open question, since it could be more time consuming to execute than the crude frequency simulator for the same number of iterations.

If $\mathrm{p}(\boldsymbol{x})$ is multivariate normal, as in the multinomial probit example, then the distributions in question are truncated univariate conditional normals, whose parameters are straightforward to derive [see Keane (1990, 1993, 1994), Geweke (1991), Borsch-Supan and Hajivassiliou (1993)]. Drawing from these distributions is also straightforward, although it is important not to use naive acceptance sampling from the unconstrained normal distribution [Geweke (1991)]. Further discussion is provided by Hajivassiliou, McFadden and Ruud (1996) and public domain software is available at http//:www.econ.umn.edu/˜bacc.

2.2. The Gibbs sampler

The Gibbs sampler is an algorithm that has been used with noted success in many econometric models. It is one example of a wider class of procedures known as *Markov chain Monte Carlo* (MCMC). MCMC constructs a Markov chain whose unique invariant distribution corresponds to the p.d.f. of the problem at hand. In the context of Equation (2.1) the invariant distribution has density $p(\boldsymbol{x})$ with respect to Lebesgue measure. In much of what follows we shall consider the more general problem arising in Bayesian inference, of finding a Markov chain with state space Θ_A and a unique invariant distribution with density $p(\theta_A|\boldsymbol{y}^o, A)$ with respect to a measure $\mathrm{d}\nu(\theta_A)$. The generic problem is then to approximate $\mathrm{E}[g(\boldsymbol{y}^o, \theta_A)|\boldsymbol{y}^o, A]$, where $g(\boldsymbol{y}, \theta_A)$ is a (possibly random) function with the property $\mathrm{E}[g(\boldsymbol{y}, \theta_A)|\boldsymbol{y}^o, \theta_A] = \mathrm{E}[h(\omega)|\boldsymbol{y}^o, A]$. Following an initial transient or *burn-in* phase, simulated values from the chain are used to approximate $\mathrm{E}[g(\boldsymbol{y}^o, \theta)|\boldsymbol{y}^o, A]$. In this generalization θ_A represents all of the unknown features of the model. This may include latent variables as well as parameters per se. In the rest of Section 2 we dispense with the "A" subscript on θ_A to reduce notational clutter.

Markov chain methods have a history in mathematical physics dating back to the algorithm of Metropolis et al. (1953). This method, which is described in Hammersly and Handscomb (1964, Section 9.3) and Ripley (1987, Section 4.7), was generalized by Hastings (1970), who focused on statistical problems, and was further explored by Peskun (1973). A version particularly suited to image reconstruction and problems in spatial statistics was introduced by Geman and Geman (1984). This was subsequently shown to have great potential for Bayesian computation by Gelfand and Smith (1990). Their work, combined with data augmentation methods [Tanner and Wong (1987)], has proven very successful in the treatment of latent variables in econometrics. Since 1990 application of MCMC methods has grown rapidly [Chib and Greenberg (1996)].

This section and the next concentrate on a heuristic development of two widely used variants of these methods, the Gibbs sampler and the Hastings–Metropolis algorithm. The general theory of convergence is taken up in Section 2.4. Section 2.5 provides a useful hybrid of the Gibbs and Hastings–Metropolis algorithms. Section 2.6 turns to the assessment of numerical accuracy.

The Gibbs sampler begins with a partition, or *blocking*, of $\theta, \theta' = (\theta'_{(1)}, \ldots, \theta'_{(B)})$. In applications, the blocking is chosen so that it is possible to draw from each of the conditional p.d.f.'s, $p(\theta_{(b)}|\boldsymbol{y}^o, \theta_{(a)}(a < b), \theta_{(a)}(a > b), A)$. This blocking can arise naturally, if the prior distributions for the $\theta_{(b)}$ are independent and each is conditionally conjugate. To motivate the key idea underlying the Gibbs sampler suppose – contrary to fact – that there existed a single drawing $\theta^{(0)}, \theta'^{(0)} = \left(\theta'^{(0)}_{(1)}, \ldots, \theta'^{(0)}_{(B)}\right)$, from $p(\theta|\boldsymbol{y}^o, A)$. Successively make drawings from the conditional distributions as follows:

$$
\begin{aligned}
&\theta^{(1)}_{(1)} \sim p\left(\cdot|\boldsymbol{y}^o, \theta^{(0)}_{(2)}, \ldots, \theta^{(0)}_{(B)}, A\right),\\
&\theta^{(1)}_{(2)} \sim p\left(\cdot|\boldsymbol{y}^o, \theta^{(1)}_{(1)}, \theta^{(0)}_{(3)}, \ldots, \theta^{(0)}_{(B)}, A\right),\\
&\quad\ldots\\
&\theta^{(1)}_{(b)} \sim p\left(\cdot|\boldsymbol{y}^o, \theta^{(1)}_{(1)}, \ldots, \theta^{(1)}_{(b-1)}, \theta^{(0)}_{(b+1)}, \ldots, \theta^{(0)}_{(B)}, A\right),\\
&\quad\ldots\\
&\theta^{(1)}_{(B)} \sim p\left(\cdot|\boldsymbol{y}^o, \theta^{(1)}_{(1)}, \ldots, \theta^{(1)}_{(B-1)}, A\right).
\end{aligned}
\tag{2.6}
$$

This defines a transition process from $\theta'^{(0)}$ to $\theta'^{(1)} = (\theta'^{(1)}_{(1)}, \ldots, \theta'^{(1)}_{(B)})$. Since $\theta^{(0)} \sim p(\theta|\boldsymbol{y}^o, A)$, $(\theta^{(1)}_{(1)}, \ldots, \theta^{(1)}_{(b-1)}, \theta^{(1)}_{(b)}, \theta^{(0)}_{(b+1)}, \ldots, \theta^{(0)}_{(B)}) \sim p(\theta|\boldsymbol{y}^o, A)$ at each step in Equation (2.6) by definition of the conditional density. In particular, $\theta^{(1)} \sim p(\theta|\boldsymbol{y}^o, A)$.

Iteration of this algorithm produces a sequence $\theta^{(0)}, \theta^{(1)}, \ldots, \theta^{(m)}, \ldots$ which is a realization of a Markov chain with probability density function kernel for the transition from point $\theta^{(m)}$ to point $\theta^{(m+1)}$ given by

$$K_G\left(\theta^{(m)}, \theta^{(m+1)}\right) = \prod_{b=1}^{B} p\left[\theta^{(m+1)}_{(b)} | \boldsymbol{y}^o, \theta^{(m)}_{(a)}(a > b), \theta^{(m+1)}_{(a)}(a < b), A\right]. \qquad (2.7)$$

Any single iterate $\theta^{(m)}$ retains the property that it is drawn from the posterior distribution. For the Gibbs sampler to be practical, it is essential that the blocking be chosen in such a way that one can make the drawings in an efficient manner. In econometrics the blocking is often natural and the conditional distributions familiar.

Of course, if it really were possible to make an initial draw from the posterior distribution, then simple frequency simulation would also be possible. An important remaining task is to elucidate conditions for the distribution of $\theta^{(m)}$ to converge to the posterior for any $\theta^{(0)} \in \Theta$. This is not trivial, because even if $\theta^{(0)}$ were drawn from $p(\theta|\boldsymbol{y}^o, A)$, the argument just given establishes only that any single $\theta^{(m)}$ is also drawn from the posterior distribution. It does not establish that a single sequence $\left\{\theta^{(m)}\right\}_{m=1}^{\infty}$ is representative of the posterior distribution. For example, if Θ consists of two disjoint subsets Θ_1 and Θ_2 with $\theta_1 > \theta_2 \;\forall\; \theta_j \in \Theta_j$, then a Gibbs sampler that begins in Θ_1 may never visit Θ_2 and vice versa. This situation clearly does not arise in the Gibbs samplers for the standard linear and probit models just described, but evidently a careful development of conditions under which $\left\{\theta^{(m)}\right\}$ converges in distribution to the posterior distribution is needed. We outline these developments in Section 2.4.

Notice that the Gibbs sampler is a method for finding a fixed, or invariant, distribution corresponding to a well-defined iteration of distributions. It is therefore analogous to other fixed-point algorithms for solving deterministic systems, such as the computation of equilibrium. In both cases, the conditions under which the fixed point is unique, and the conditions under which convergence to this fixed point is known to occur, are important. We take up these theoretical points below in Sections 2.4 and 2.6, and further practical guidance is provided in Geweke (1992), Gelman, Carlin, Stern and Rubin (1995, Chapter 11), and in software available at http://www.econ.umn.edu/˜bacc. First we turn to a related algorithm that often complements the Gibbs sampler.

2.3. The Hastings–Metropolis algorithm

The Hastings–Metropolis algorithm begins with an arbitrary transition probability density function $q(\theta_1, \theta_2)$ indexed by $\theta_1 \in \Theta$ and with density argument $\theta_2 \in \Theta$, and with an arbitrary starting value $\theta^{(0)} \in \Theta$. The random vector θ^* generated from

$q(\theta^{(m)}, \theta^{*})$ is a candidate value for $\theta^{(m+1)}$. The algorithm actually sets $\theta^{(m+1)} = \theta^{*}$ with probability

$$\alpha(\theta^{(m)}, \theta^{*}) = \min\left\{\frac{p(\theta^{*}|\boldsymbol{y}^{o}, A)\, q(\theta^{*}, \theta^{(m)})}{p(\theta^{(m)}|\boldsymbol{y}^{o}, A) q(\theta^{(m)}, \theta^{*})}, 1\right\} = \min\left\{\frac{p(\theta^{*}|\boldsymbol{y}^{o}, A)/q(\theta^{(m)}, \theta^{*})}{p(\theta^{(m)}|\boldsymbol{y}^{o}, A)/q(\theta^{*}, \theta^{(m)})}, 1\right\}; \tag{2.8}$$

otherwise, the algorithm sets $\theta^{(m+1)} = \theta^{(m)}$. This defines a Markov chain with a generally mixed continuous-discrete transition probability from $\theta^{(m)}$ to $\theta^{(m+1)}$ given by

$$K_H\left(\theta^{(m)}, \theta^{(m+1)}\right) = \begin{cases} q\left(\theta^{(m)}, \theta^{(m+1)}\right) \alpha\left(\theta^{(m)}, \theta^{(m+1)}\right) & \text{if} \quad \theta^{(m+1)} \neq \theta^{(m)}, \\ 1 - \int_{\Theta} q\left(\theta^{(m)}, \theta\right) \alpha\left(\theta^{(m)}, \theta\right) \mathrm{d}\nu(\theta) & \text{if} \quad \theta^{(m+1)} = \theta^{(m)}. \end{cases}$$

This form of the algorithm is due to Hastings (1970). The Metropolis et al. (1953) form took $q\left(\theta^{(m)}, \theta^{*}\right) = q\left(\theta^{*}, \theta^{(m)}\right)$.

There is a simple two-step argument that motivates the convergence of the sequence $\left\{\theta^{(m)}\right\}$ generated by the Hastings–Metropolis algorithm to the posterior. [This approach is due to Chib and Greenberg (1995a)]. First, observe that if the transition probability function $p\left(\theta^{(m)}, \theta^{(m+1)}\right)$ satisfies the *reversibility condition*

$$p\left(\theta^{(m)}\right) p\left(\theta^{(m)}, \theta^{(m+1)}\right) = p\left(\theta^{(m+1)}\right) p\left(\theta^{(m+1)}, \theta^{(m)}\right), \tag{2.9}$$

for stated $p(\cdot)$, then it has $p(\cdot)$ as an invariant distribution. To see this, note that if Equation (2.9) holds then

$$\begin{aligned} \int_{\Theta} p\left(\theta^{(m)}\right) p\left(\theta^{(m)}, \theta^{(m+1)}\right) \mathrm{d}\nu\left(\theta^{(m)}\right) &= \int_{\Theta} p\left(\theta^{(m+1)}\right) p\left(\theta^{(m+1)}, \theta^{(m)}\right) \mathrm{d}\nu\left(\theta^{(m)}\right) \\ &= p\left(\theta^{(m+1)}\right) \int_{\Theta} p\left(\theta^{(m+1)}, \theta^{(m)}\right) \mathrm{d}\nu\left(\theta^{(m)}\right) = p\left(\theta^{(m+1)}\right). \end{aligned}$$

For $\theta^{(m+1)} = \theta^{(m)}$, Equation (2.9) is satisfied trivially. For $\theta^{(m+1)} \neq \theta^{(m)}$, suppose without loss of generality that $p\left(\theta^{(m+1)}\right)/q\left(\theta^{(m)}, \theta^{(m+1)}\right) > p\left(\theta^{(m)}\right)/q\left(\theta^{(m+1)}, \theta^{(m)}\right)$. Then

$$p\left(\theta^{(m)}, \theta^{(m+1)}\right) = q\left(\theta^{(m)}, \theta^{(m+1)}\right),$$

and

$$\begin{aligned} p\left(\theta^{(m+1)}, \theta^{(m)}\right) &= q\left(\theta^{(m+1)}, \theta^{(m)}\right) \cdot \frac{p\left(\theta^{(m)}\right)/q\left(\theta^{(m+1)}, \theta^{(m)}\right)}{p\left(\theta^{(m+1)}\right)/q\left(\theta^{(m)}, \theta^{(m+1)}\right)} \\ &= p\left(\theta^{(m)}\right) q\left(\theta^{(m)}, \theta^{(m+1)}\right)/p\left(\theta^{(m+1)}\right), \end{aligned}$$

whence Equation (2.9) is satisfied.

In implementing the Hastings–Metropolis algorithm the transition probability density function must share two important properties. First, it must be possible to generate θ^* efficiently from $q\left(\theta^{(m)}, \theta^*\right)$. A second key characteristic of a satisfactory transition process is that the unconditional acceptance rate not be so low that the time required to generate a sufficient number of distinct $\theta^{(m)}$ is too great.

2.4. Some Markov chain Monte Carlo theory

Much of the treatment here draws heavily on the work of Tierney (1994), who first used the theory of general state space Markov chains to demonstrate convergence, and Roberts and Smith (1994), who elucidated sufficient conditions for convergence that turn out to be applicable in a wide variety of problems in econometrics.

Let $\left\{\theta^{(m)}\right\}_{m=0}^{\infty}$ be a Markov chain defined on $\Theta \subseteq \Re^k$ with transition density K: $\Theta \times \Theta \to \Re^+$ such that, for all ν-measurable $\Theta_0 \subseteq \Theta$,

$$P\left(\theta^{(m)} \in \Theta_0 | \theta^{(m-1)}\right) = \int_{\Theta_0} K\left(\theta^{(m-1)}, \theta\right) \, \mathrm{d}\nu(\theta) + r\left(\theta^{(m-1)}\right) \chi_{\Theta_0}\left(\theta^{(m-1)}\right),$$
$$\text{where} \quad r\left(\theta^{(m-1)}\right) = 1 - \int_{\Theta} K\left(\theta^{(m-1)}, \theta\right) \, \mathrm{d}\nu(\theta).$$

The transition density K is substochastic: it defines only the distribution of accepted candidates. Assume that K has no absorbing states, so that $r(\theta) < 1 \ \forall \, \theta \in \Theta$. The corresponding substochastic kernel over m steps is then defined iteratively,

$$K^{(m)}\left(\theta^{(0)}, \theta^{(m)}\right) = \int_{\Theta} K^{(m-1)}\left(\theta^{(0)}, \theta\right) K\left(\theta, \theta^{(m)}\right) \, \mathrm{d}\nu(\theta)$$
$$+ K^{(m-1)}\left(\theta^{(0)}, \theta^{(m)}\right) r\left(\theta^{(m)}\right) + \left[r\left(\theta^{(0)}\right)\right]^{m-1} K\left(\theta^{(0)}, \theta^{(m)}\right).$$

This describes all m-step transitions that involve at least one accepted move. As a function of $\theta^{(m)}$ it is the p.d.f. with respect to ν of $\theta^{(m)}$, excluding realizations with $\theta^{(n)} = \theta^{(0)} \, \forall \, n = 1, \ldots, m$. For any ν-measurable Θ_0 let $P^{(m)}\left(\theta^{(0)}, \Theta_0\right)$ denote the m'th iterate of P,

$$P^{(m)}\left(\theta^{(0)}, \Theta_0\right) = \int_{\Theta_0} K^{(m)}\left(\theta^{(0)}, \theta\right) \, \mathrm{d}\nu(\theta) + \left[r\left(\theta^{(0)}\right)\right]^{m} \chi_{\Theta_0}\left(\theta^{(0)}\right).$$

An invariant distribution of the transition density K is a function $p(\theta)$ that satisfies

$$P(\Theta_0) = \int_{\Theta_0} p(\theta) \, \mathrm{d}\nu(\theta) = \int_{\Theta} P\left(\theta^{(m)} \in \Theta_0 | \theta^{(m-1)} = \theta\right) p(\theta) \, \mathrm{d}\nu(\theta)$$
$$= \int_{\Theta} \left\{ \int_{\Theta_0} K(\theta, \theta^*) \, \mathrm{d}\nu(\theta^*) + r(\theta) \chi_{\Theta_0}(\theta) \right\} p(\theta) \, \mathrm{d}\nu(\theta),$$

for all ν-measurable Θ_0. Let $\Theta^* = \{\theta \in \Theta : p(\theta) > 0\}$. The density K is *p-irreducible* if for all $\theta^{(0)} \in \Theta^*, P(\Theta_0) > 0$ implies that $P^{(m)}\left(\theta^{(0)}, \Theta_0\right) > 0$ for some $m \geqslant 1$.

The transition density K is *aperiodic* if there exists no ν-measurable partition $\Theta = \bigcup_{s=0}^{r-1} \tilde{\Theta}_s (r \geqslant 2)$ such that

$$P\left(\theta^{(m)} \in \tilde{\Theta}_{m \bmod (r)} | \theta^{(0)} \in \tilde{\Theta}_0\right) = 1 \; \forall \, m.$$

It is *Harris recurrent* if $P\left[\theta^{(m)} \in \Theta_0 \text{ i.o. } |\theta^{(0)}\right] = 1$ for all ν-measurable Θ_0 with $\int_{\Theta_0} p(\theta) \, \mathrm{d}\nu(\theta) > 0$ and all $\theta^{(0)} \in \Theta$.[2] It follows directly that if a kernel is Harris recurrent, then it is p-irreducible. A kernel whose invariant distribution is proper, and that is both aperiodic and Harris recurrent, is *ergodic* by definition [Tierney (1994, pp. 1712–1713)].

A useful metric in what follows is the total variation norm for signed and bounded measures μ defined over the field of all ν-measurable sets S_ν on Θ: $|\mu| = \sup_{\Theta_0 \in S_\nu} \mu(\Theta_0) - \inf_{\Theta_0 \in S_\nu} \mu(\Theta_0)$.

Theorem 2.4.1. Convergence of continuous state Markov chains. *Suppose $p(\theta|\boldsymbol{y}^o, A)$ is an invariant distribution of the transition density $K(\theta, \theta^*)$.*

(A) If K is $p(\theta|\boldsymbol{y}^o, A)$-irreducible, then $p(\theta|\boldsymbol{y}^o, A)$ is the unique invariant distribution.

(B) If K is $p(\theta|\boldsymbol{y}^o, A)$-irreducible and aperiodic, then except possibly for $\theta^{(0)}$ in a set of posterior probability 0, $\left|P^{(m)}\left(\theta^{(0)}, \cdot\right) - P(\cdot|\boldsymbol{y}^o, A)\right| \to 0$. If K is ergodic (that is, it is also Harris recurrent) then this occurs for all $\theta^{(0)}$.

(C) If K is ergodic with invariant distribution $p(\theta|\boldsymbol{y}^o, A)$, then for all $g(\boldsymbol{y}^o, \theta)$ absolutely integrable with respect to $p(\theta|\boldsymbol{y}^o, A)$ and for all $\theta^{(0)} \in \Theta$,

$$M^{-1} \sum_{m=1}^{M} g\left(\boldsymbol{y}^o, \theta^{(m)}\right) \xrightarrow{\text{a.s.}} \int_{\Theta} g(\boldsymbol{y}^o, \theta) p(\theta|\boldsymbol{y}^o, \theta) \, \mathrm{d}\nu(\theta).$$

Proof: (A) and (B) follow immediately from Theorem 1, and (C) from Theorem 3 in Tierney (1994). □

For the Gibbs sampling algorithm we argued informally in Section 2.2 that $p(\theta|\boldsymbol{y}^o, A)$ is an invariant distribution. More formally, from Equation (2.7) we have for the blocking $\theta' = \left(\theta'_{(1)}, \theta'_{(2)}\right)$,

$$\begin{aligned}
&\int_{\Theta} K_G(\theta, \theta^*) p(\theta|\boldsymbol{y}^o, A) \, \mathrm{d}\nu(\theta) \\
&\quad = \int_{\Theta} p\left(\theta^*_{(1)}|\boldsymbol{y}^o, \theta_{(2)}, A\right) p\left(\theta^*_{(2)}|\boldsymbol{y}^o, \theta^*_{(1)}, A\right) p(\theta|\boldsymbol{y}^o, A) \, \mathrm{d}\nu(\theta) \\
&\quad = p\left(\theta^*_{(2)}|\boldsymbol{y}^o, \theta^*_{(1)}, A\right) \int_{\Theta} p\left(\theta^*_{(1)}|\boldsymbol{y}^o, \theta_{(2)}, A\right) p(\theta|\boldsymbol{y}^o, A) \, \mathrm{d}\nu(\theta) \\
&\quad = p\left(\theta^*_{(2)}|\boldsymbol{y}^o, \theta^*_{(1)}, A\right) p\left(\theta^*_{(1)}|\boldsymbol{y}^o, A\right) = p(\theta^*|\boldsymbol{y}^o, A).
\end{aligned}$$

[2] The expression "i.o." in $P[\theta^{(m)} \in \Theta_0 \text{ i.o. } |\theta^{(0)}]$ means "infinitely often". The condition is that $\lim_{M \to \infty} P[\sum_{m=1}^{M} \chi_{\tilde{\Theta}}(\theta^{(m)}) \leqslant L] = 0 \; \forall \, L$.

The general result for more than two blocks follows by induction. Thus, it is the uniqueness of the invariant state that is at issue in establishing convergence of the Gibbs sampler. The following result is immediate and is often easy to apply.

Corollary 2.4.2. A first sufficient condition for convergence of the Gibbs sampler. Suppose that for every point $\theta^* \in \Theta$ and every $\Theta_0 \subseteq \Theta$ with the property $P(\theta \in \Theta_0|\boldsymbol{y}^o, A) > 0$, it is the case that $P_G\left(\theta^{(m+1)} \in \Theta_0|\boldsymbol{y}^o, \theta^{(m)} = \theta^*, A\right) > 0$, where $P_G(\cdot)$ is the probability measure induced by the Gibbs sampler. Then the Gibbs transition kernel is ergodic.

Proof: The conditions ensure that P_G is aperiodic and absolutely continuous with respect to $p(\theta|\boldsymbol{y}^o, A)$. The result follows from Corollary 1 of Tierney (1994). □

A complement to Corollary 2.4.2 is provided by Roberts and Smith (1994).

Theorem 2.4.3. A second sufficient condition for convergence of the Gibbs sampler. *Suppose that $p(\theta|\boldsymbol{y}^o, A)$ is lower semicontinuous*[3] *at 0 and $\int_{\Theta^{(b)}} p(\theta|\boldsymbol{y}^o)\, \mathrm{d}\nu\left(\theta^{(b)}\right)$ is locally bounded ($b = 1, \ldots, B$). Suppose also that Θ is connected. Then the Gibbs transition kernel is ergodic.* ■

Tierney (1994) discusses weaker conditions for convergence of the Gibbs sampler. However, the conditions of Corollary 2.4.2 or Theorem 2.4.3 are satisfied for a very wide range of problems in econometrics and are easier to verify.

Tierney (1994) and Roberts and Smith (1994) show that the convergence properties of the Hastings–Metropolis algorithm are inherited from those of $q(\theta, \theta^*)$: if q is aperiodic and $p(\theta|\boldsymbol{y}^o, A)$-irreducible, then so is the Hastings–Metropolis algorithm. This feature leads to a sufficient condition for convergence analogous to Corollary 2.4.2.

Theorem 2.4.4. A first sufficient condition for convergence of the Hastings–Metropolis algorithm. *Suppose that for every point $\theta^* \in \Theta$ and every $\Theta_0 \subseteq \Theta$ with the property $P(\theta \in \Theta_0|\boldsymbol{y}^o, A) > 0$, it is the case that $\int_{\Theta_0} q(\theta, \theta^*)\, \alpha(\theta, \theta^*)\, \mathrm{d}\nu(\theta^*) + r(\theta)\chi_{\Theta_0}(\theta) > 0$. Then the Hastings–Metropolis density $K(\theta, \theta^*) = q(\theta, \theta^*)\, \alpha(\theta, \theta^*)$ is ergodic.*

Proof: The conditions ensure that the transition kernel is aperiodic and $p(\theta^*|\boldsymbol{y}^o, A)$-irreducible. Thus, by Corollary 2 of Tierney (1994), the Hastings–Metropolis density is Harris recurrent. Since the kernel is both aperiodic and Harris recurrent, it is ergodic. □

A complementary sufficient condition for convergence of Hastings–Metropolis chains is provided by the following result, which is analogous to Theorem 2.4.3 for the Gibbs sampler.

[3] A function $h(\boldsymbol{x})$ is lower semicontinuous at 0 if, for all $\boldsymbol{x}$ with $h(\boldsymbol{x}) > 0$, there exists an open neighborhood $N_x \supset \boldsymbol{x}$ and $\varepsilon > 0$ such that for all $\boldsymbol{y} \subset N_x, h(\boldsymbol{y}) \geqslant \varepsilon > 0$.

Theorem 2.4.5. A second sufficient condition for convergence of the Hastings–Metropolis algorithm. *Suppose that for every $\theta \in \Theta$, $p(\theta|\boldsymbol{y}^o, A) > 0$, and for all pairs $\left(\theta^{(m)}, \theta^{(m+1)}\right) \in \Theta \times \Theta$, $p\left(\theta^{(m)}|\boldsymbol{y}^o, A\right)$ and $q\left(\theta^{(m)}, \theta^{(m+1)}\right)$ are positive and continuous. Then the Hastings–Metropolis kernel K_H is ergodic.*

Proof: See Chib and Greenberg (1995a) or Mengersen and Tweedie (1996). □

Once again, the conditions are sufficient but not necessary, but weaker conditions are typically more difficult to verify. On weaker conditions, see Tierney (1994).

2.5. Metropolis within Gibbs

There are many variations on these methods, and alone or in combination with each other they provide a powerful source of flexibility that can be drawn upon in constructing posterior simulators. Here we briefly review one, which has been quite useful in econometrics and will be used subsequently in Section 6 of this chapter. Further discussion can be found in Tierney (1994), Gelman et al. (1995) and Geweke (1999).

The Metropolis within Gibbs algorithm [Zeger and Karim (1991), Chib and Greenberg (1996)] provides a neat solution to the problem of a block $\theta_{(b)}$ in a Gibbs sampler, in which it is difficult to draw directly from $p\left(\theta_{(b)}|\boldsymbol{y}^o, \theta_{(a)}(a < b), \theta_{(a)}(a > b), A\right)$. In a two-block Gibbs sampler, suppose that it is straightforward to sample from $p\left(\theta_{(1)}|\boldsymbol{y}^o, \theta_{(2)}, A\right)$, but the distribution corresponding to $p\left(\theta_{(2)}|\boldsymbol{y}^o, \theta_{(1)}, A\right)$ is intractable. The Hastings–Metropolis algorithm can be used in these circumstances, and it often provides an efficient solution to the problem. In what has become known as the Metropolis-within-Gibbs procedure, at the $(m+1)$'th iteration first draw $\theta^*_{(2)}$ from a proposal density $q\left(\theta^{(m)}_{(2)}, \theta^*_{(2)}|\theta^{(m+1)}_{(1)}\right)$. Accept this draw with probability

$$\min\left\{\frac{p\left(\theta^{(m+1)}_{(1)}, \theta^*_{(2)}|\boldsymbol{y}^o, A\right)/q\left(\theta^{(m)}_{(2)}, \theta^*_{(2)}|\theta^{(m+1)}_{(1)}\right)}{p\left(\theta^{(m+1)}_{(1)}, \theta^{(m)}_{(2)}|\boldsymbol{y}^o, A\right)/q\left(\theta^*_{(2)}, \theta^{(m)}_{(2)}|\theta^{(m+1)}_{(1)}\right)}, 1\right\}.$$

If $\theta^*_{(2)}$ is accepted then $\theta^{(m+1)}_{(2)} = \theta^*_{(2)}$, and if not then $\theta^{(m+1)}_{(2)} = \theta^{(m)}_{(2)}$. The extension of this procedure to multi-block Gibbs samplers, with a Hastings–Metropolis algorithm used at some (or even all) of the blocks is clear. For further discussion see Chib and Greenberg (1995a), and for a proof that the posterior distribution is an invariant state of this Markov chain see Chib and Greenberg (1996).

2.6. Assessing numerical accuracy in Markov chain Monte Carlo

In any practical application one is concerned with the discrepancy $\bar{g}_M - \bar{g}$. A leading analytical tool for assessing this discrepancy is a central limit theorem, if one can be obtained. This was accomplished in Section 2.1 for i.i.d. sampling from the posterior distribution. The assumption of independence, key to those results, does not apply in Markov chain Monte Carlo. The weaker assumption of uniform ergodicity yields a central limit theorem, however. Let $P^{(m)}\left(\theta^{(0)}, \Theta_0\right)$ denote $P\left(\theta^{(m)} \in \Theta_0 | \theta^{(0)}\right)$ for any $\theta^{(0)} \in \Theta$ and for any $\Theta_0 \subseteq \Theta$ for which $P(\theta \in \Theta_0 | \boldsymbol{y}^o, A)$ is defined. The Markov chain is *uniformly ergodic* if $\sup_{\theta \in \Theta} \left|P^{(m)}(\theta, \cdot) - P(\cdot | \boldsymbol{y}^o, A)\right| \leqslant Mr^m$ for some $M > 0$ and some positive $r < 1$. Tierney (1994, p. 1714) provides results that are quite useful in establishing uniform ergodicity. The main result is the following.

Theorem 2.6.1. A central limit theorem for Markov chain Monte Carlo. *Suppose* $\left\{\theta^{(m)}\right\}$ *is uniformly ergodic with equilibrium distribution* $p(\theta | \boldsymbol{y}^o, A)$. *Suppose further that* $\mathrm{E}\left[g(\boldsymbol{y}^o, \theta) \,|\, \boldsymbol{y}^o, A\right] = \bar{g}$ *and* $var\left[g(\boldsymbol{y}^o, \theta) \,|\, \boldsymbol{y}^o, A\right]$ *exist and are finite, and let* $\bar{g}_M = M^{-1} \sum_{m=1}^{M} g\left(\boldsymbol{y}^o, \theta^{(m)}\right)$. *Then there exists finite* σ^2 *such that*

$$M^{1/2}(\bar{g}_M - \bar{g}) \xrightarrow{d} N\left(0, \sigma^2\right). \tag{2.10}$$

Proof: Tierney (1994, Theorem 5), attributed to Cogburn (1972, Corollary 4.2(ii)). □

A key difficulty in implementing this result is that useful conditions sufficient for approximation of the unknown constant σ^2 have not yet been developed. That is, there is no $\hat{\sigma}_M^2$ for which $\hat{\sigma}_M^2 \to \sigma^2$ as there is for independence and importance sampling. A second difficulty is assessing the sensitivity of $\theta^{(m)}$ to the initial condition $\theta^{(0)}$. For example, if the posterior density is multimodal the Markov chain may be nearly reducible. Assessing convergence in such situations is clearly nontrivial.

There is an extensive literature on this problem. A good introduction is provided by the papers of Gelman and Rubin (1992) and Geyer (1992) and their discussants. Geweke (1992) developed a consistent estimator of σ^2 in Equation (2.10), under the strong condition that conventional time series mixing conditions [for example Hannan (1970, pp. 207–210) apply to $\left\{\theta^{(m)}\right\}$. There is no analytical foundation for this assumption, but these methods are now widely used and have proven reliable in the sense that they predict well the behavior of the Markov chain when it is restarted with a new initial condition, in econometric models.

It is often useful to compare the estimate $\hat{\sigma}^2$ of σ^2 in Equation (2.10) with $\mathrm{var}\left[g(\boldsymbol{y}^o, \theta) \,|\, \boldsymbol{y}^o, A\right]$. Recall that the latter would replace σ^2, if i.i.d. sampling directly from the posterior distribution were possible and employed. Since $\mathrm{var}\left[g(\boldsymbol{y}^o, \theta) \,|\, \boldsymbol{y}^o, A\right]$ can be approximated from $\left\{\theta^{(m)}\right\}_{m=1}^{M}$, it is possible to approximate the *relative numerical efficiency* $\sigma^2/\mathrm{var}\left[g(\boldsymbol{y}^o, \theta) \,|\, \boldsymbol{y}^o, A\right]$. This quantity, introduced in Geweke (1989), indicates the number of simulations that would have been required under i.i.d. sampling from the posterior distribution, relative to the number required in the sampling scheme at hand. In MCMC algorithms low values of RNE indicate strong serial correlation in the Markov chain.

In practice, some robustness to initial conditions is achieved by discarding initial iterations: 10% to 20% is common. By drawing $\theta^{(0)}$ from the prior distribution, using a random number generator with a fresh seed each time, several runs may provide some indication of whether the results are sensitive to initial conditions as they might be, for example, given near-reducibility of the kind that may arise from severe multimodality. A formal test for sensitivity to initial conditions was developed by Gelman and Rubin (1992) and is described in Section 3.8. For other tests for sensitivity to initial conditions see Geweke (1992) and Zellner and Min (1995).

3. Approximate solution of discrete dynamic optimization problems

Consider an agent who controls the decision d_t at each time $t = 1, \ldots, T < \infty$, where $d_t \in \{1, \ldots, J\}$. The agent's state is characterized by the $S \times 1$ vector I_t. The state evolves according to the p.d.f. $p(I_{t+1}|I_t, d_t)$. It also is sometimes useful to express (equivalently) the evolution of the state variables using the random function $I_{t+1} = M(I_t, d_t)$. The current period payoff to decision d_t in state I_t is $u_{d_t}(I_t)$. The agent's problem is:

$$\max_{d_t} \left\{ u_{d_t}(I_t) + E\left[\sum_{\tau=t+1}^{T} \delta^{\tau-t} u_{d_\tau}(I_\tau) \mid I_t, d_t \right] \right\}, \tag{3.1}$$

where δ is the discount factor. The operator $E[\cdot|\cdot]$ denotes the agent's subjective expectation over future states, choices and payoffs, given the information set (I_t, d_t). The agent faces a dynamic optimization problem if this expectation depends on d_t in a nontrivial way. The alternative specific value function associated with the choice $d_t = j$ in state I_t is:

$$V_j(I_t) = u_j(I_t) + E\left[\sum_{\tau=t+1}^{T} \delta^{\tau-t} u_{d_\tau}(I_\tau) \mid I_t, j \right] \quad (t \leqslant T). \tag{3.2}$$

It will be useful in much of the discussion to refer to the expectation term in Equation (3.2) as the "future component" of the value function, and to denote it by $F_{t+1}(I_t, d_t)$. This future component, divided by δ, is often referred to as the "Emax" function, for reasons that become obvious if we rewrite it in the form:

$$\begin{aligned} F_{t+1}(I_t, d_t) &= \delta E\left[\max_{d_{t+1}} \left\{ u_{d_{t+1}}(I_{t+1}) + E \sum_{\tau=t+2}^{T} \delta^{\tau-t-1} u_{d_\tau}(I_\tau) \right\} \mid I_t, d_t \right] \\ &= \delta E\left[\max_{d_{t+1}} \{ V_{d_{t+1}}(I_{t+1}) \} \mid I_t, d_t \right]. \end{aligned}$$

Solution of a discrete dynamic optimization problem requires that the Emax functions be evaluated at every possible (I_t, d_t) combination for $t = 1, \ldots, T - 1$. Note that at the terminal period T there is, by definition, no future component of the value functions, and we simply have:

$$V_j(I_T) = u_j(I_T).$$

Hence,

$$F_T(I_{T-1}, d_{T-1}) = \delta E\left[\max_{d_T}\{u_{d_T}(I_T)\} \mid I_{T-1}, d_{T-1}\right]. \tag{3.3}$$

Assume for now that expectations are formed rationally, in the sense that the operator $E[\cdot|\cdot]$ is the mathematical expectation operator. Then, the Emax function in Equation (3.3) is an integral over the stochastic terms that enter either the payoff functions $\{u_j(I_T)\}_{j=1}^{J}$ or the function M that maps (I_{T-1}, d_{T-1}) into I_T. Since there is no future component beyond T, if the functional forms for u and M are known, it is (in principle) straightforward to evaluate such integrals and construct the $F_T(I_{T-1}, d_{T-1})$ function associated with every possible combination (I_{T-1}, d_{T-1}). For this reason, the standard solution method for finite horizon discrete dynamic optimization problems such as Equation (3.1) is to "*backsolve*". Note that at time $T-1$ the alternative specific value functions are given by

$$V_j(I_{T-1}) = u_j(I_{T-1}) + F_T(I_{T-1}, j).$$

Since the $F_T(I_{T-1},j)$ were calculated in the first stage of the backsolving process, it is (in principle) straightforward to construct the functions

$$\begin{aligned} F_{T-1}(I_{T-2}, d_{T-2}) &= \delta E\left[\max_{d_{T-1}}\{V_{d_{T-1}}(I_{T-1})\} \mid I_{T-2}, d_{T-2}\right] \\ &= \delta E\left[\max_{d_{T-1}}\{u_{d_{T-1}}(I_{T-1}) + F_T(I_{T-1}, d_{T-1})\} \mid I_{T-2}, d_{T-2}\right]. \end{aligned} \tag{3.4}$$

Note that the calculations involved in taking the expectation here are no more complex than those involved in Equation (3.3) since the $F_T(I_{T-1}, d_{T-1})$ are known constants at this stage of the backsolving process. The Emax functions in Equation (3.4) are again simply integrals over the stochastic terms that enter the payoff functions $\{u_j(I_{T-1})\}_{j=1}^{J}$ or the function M mapping (I_{T-2}, d_{T-2}) into I_{T-1}. After evaluating these integrals for every possible (I_{T-2}, d_{T-2}) combination we move back to period $T-2$, and so on until the backsolving process is complete. The fact that a T period problem like Equation (3.1) can be cast as a series of two period problems in this way is referred to as the Bellman (1957) principle.

There are two important practical problems that often arise in this context, however:

(1) If the number of stochastic terms in the payoff functions and the law of motion M is large then the integration required to construct the Emax functions will be high dimensional.

(2) If the number of (state, decision) combinations $\left\{\{(I_t,j)\}_{j=1}^{J}\right\}_{t=1}^{T-1}$ is large then the number of integrals that must be evaluated may be large.

Problem 1 is obviously the type of high dimensional integration problem that is the focus of this chapter. However, in many contexts the problem of simulating the Emax function is not severe. Let the state vector I_t be decomposed as $\left(\bar{I}_t, \varepsilon_t\right)$, where $\bar{I}_t$ and ε_t are the elements that are nonstochastic and stochastic, respectively, from the perspective of the agent at time $t-1$. Keane and Wolpin (1994) find in a number of numerical examples that if crude frequency simulators of the form

$$M^{-1}\sum_{m=1}^{M}\sum_{j=1}^{J} V_j\left(\bar{I}_t, \varepsilon_t^m\right)\chi\left[V_j\left(\bar{I}_t, \varepsilon_t^m\right) > V_k\left(\bar{I}_t, \varepsilon_t^m\right) \quad \forall\ k \neq j\right], \tag{3.5}$$

are substituted for the exact Emax functions then deterioration of the accuracy of the solution of the optimization problem is small even with fairly small M. But such crude simulators will be highly inefficient in contexts where one or more alternatives have low probability of being realized, but where these alternatives deliver extreme (either very large or very small payoffs) when realized. In such contexts more sophisticated smooth simulators for the Emax function may be necessary[4]. Recently, Rust (1997) has advocated use of deterministic integration methods based on "low discrepancy" points rather than draws from random number generators to increase accuracy of Emax approximations. Finally, we note that a closed form for the Emax function exists in the special case noted by Rust (1987). This is when there is exactly one stochastic term entering each payoff function, those terms enter additively, and they are i.i.d. type I extreme value distributed.

Problem 2 is the main focus of the rest of this section. This problem is often referred to as the "curse of dimensionality" [see Bellman (1957)][5]. It tends to be severe when

[4] For example, a GHK type simulator of the probability that any alternative $j = 1, \ldots, J$ is optimal may be formed provided the $J-1$ conditions $V_k(\bar{I}_t, \varepsilon_t) - V_j(\bar{I}_t, \varepsilon_t) < 0\ \forall\ k \neq j$ form a sequential partition of the error space, with tractable conditional densities, so that it is possible to draw, element-by-element, vectors ε_t^m that satisfy all $J-1$ conditions. Let $\hat{p}_{\mathrm{GHK},m}(j|I_t)$ denote the GHK simulator of the probability that j is optimal, based on the single draw sequence m. Let $\varepsilon_t^{m(j)}$ denote the draw for ε_j in that draw sequence. If these quantities are obtained for $j = 1, \ldots, J$ then the (unbiased) GHK simulator for the Emax function is:

$$\hat{E}\mathrm{max}_j\left\{V_j\left(I_j\right)\right\} = M^{-1}\sum_{j=1}^{J}\sum_{m(j)=1}^{M} V_j\left(\bar{I}_t, \varepsilon_j^{m(j)}\right)\hat{p}_{\mathrm{GHK},m}(j|I_t).$$

(Note that $\varepsilon_j^{m(j)} \neq \varepsilon_j^{m(k)}$ for $k \neq j$ because different draw sequences are constructed to simulate the choice probabilities for alternatives j and k.)

[5] Formally, the curse of dimensionality refers to the *exponential* rise in computation time as the number of state and decision variables increases.

the number of state variables is large and/or some individual state variables take on a large number of values. Note, however, that in many applications it is only a subset of the complete set of state variables that is relevant in determining the computational burden involved in solving the dynamic optimization problem. This is because in many applications only a subset of the state variables enter the conditioning set in Equation (3.1) in a non-trivial way. For example, consider a simple labor supply model with human capital accumulation:

$$\begin{aligned} d_t &\in \{0,1\}, & X_{t+1} &= X_t + d_t \quad X_1 = 0, \\ \ln W_t &= \beta_0 + \beta_1 X_t + \varepsilon_t, & \varepsilon_t &\sim \text{ i.i.d. } N(0,\sigma^2), \\ I_t &= (X_t, \varepsilon_t), & u_{d_t}(I_t) &= d_t W_t + (1-d_t)\, b, \end{aligned} \tag{3.6}$$

where b is the utility from leisure, and $t = 1, \ldots, T < \infty$. Note that although the stochastic term $\varepsilon_t \in I_t$, it is not useful for forecasting payoffs at $t = t+1, \ldots, T$ because of the i.i.d. assumption. Hence, in the backsolving process it is only necessary that the Emax functions be evaluated at every possible (X_t, d_t) combination for $t = 1, \ldots, T-1$. Denote by $I_t^* \subseteq I_t$ the subset of state variables which are relevant in solution of the dynamic optimization problem, in that their values at time t influence expected payoffs in future periods.

Furthermore, it is often the case that not all possible (I_t^*, d_t) combinations need be considered in the backsolving process. For instance, in model (3.6), $F_{t+1}(X_t = a, d_t = 0)$ and $F_{t+1}(X_t = a-1, d_t = 1)$ are identical, since in both cases $X_{t+1} = a$. In the event that the number of (I_t^*, d_t) combinations that map into unique I_{t+1}^* values is finite, we denote that number by N_{t+1}^S, and let $N^S = \sum_{t=2}^T N_t^S$.

An extreme case of the curse of dimensionality arises if one or more of the state variables in I_t^* can take on an infinite number of values. Then exact solution of the discrete dynamic optimization problem is impossible, because the Emax functions must be evaluated at an infinite number of $(I_t^*, d_t)_{t=1}^{T-1}$ combinations. In such cases the available solution methods are discretization [see e.g., Santos and Vigo-Aguiar (1998)], Rust's randomization method [see Rust (1997)], and the use of functional approximations to the Emax functions [see e.g., Bellman, Kalaba and Kotkin (1963), Keane and Wolpin (1994)].

For a class of problems with continuous decision variables Santos and Vigo-Aguiar (1998) show, under the assumption that the state variables lie in a polyhedron, that as the mesh size of the grid used for discretization decreases, the approximate value function converges quadratically to the true value function, and the approximate decision rule converges linearly to the true decision rule[6]. The Santos and Vigo-Aguiar results rely on the value function being continuous in the state variables and having bounded second derivatives, as well as on the decision rule being continuous in the

[6] That is, if the mesh size is h the approximation error for the value function is bounded by Mh^2 and that for the decision rule is bounded by Nh, where M and N are positive constants.

state variables. Thus, their results are not directly applicable to models with discrete decision and/or state variables which are our main focus.

More generally, all discretization methods suffer from a curse of dimensionality if multiple state variables can take on an infinite number of values and must be discretized. Even if only one state variable must be discretized, the dimensionality problem may be severe if that variable must then be crossed with several other finite-valued state variables. For this reason, discretization is typically only a practical option in models with only one or two infinite valued state variables and few additional finite valued state variables.

Even in cases where the complete state vector $\{I_t^*\}_{t=1}^{T-1}$ takes on only a finite number of values, calculation of the Emax functions at *all* N^S of the relevant $\left\{(I_t^*,j)_{j=1}^J\right\}_{t=1}^{T-1}$ combinations is infeasible if N^S is too large. What value of N^S is "too large" depends on the speed of available computers. Also, since at each relevant (I_t^*, d_t) combination an integral must be evaluated (or simulated) to construct the $F_{t+1}(I_t^*, d_t)$ function, the maximum feasible value of N^S will be less as the time required per integration is greater.

In cases where N^S is too large to permit evaluation of the complete set of $\left\{\{F_{t+1}(I_t^*,j)\}_{j=1}^J\right\}_{t=1}^{T-1}$ functions, discretization remains an option. However, it then takes the form of adopting coarser grids for already discrete state variables. An often neglected point is that, for this procedure to be effective, it is usually the case that the decision set must also be modified accordingly. For example, in the model (3.6), assume that the state variable X_t is work experience measured in years, and $d_t = 1$ corresponds to the decision to work for one year. If we set $T = 40$, then N^S for this model is $(40 \cdot 41/2) - 1 = 819$. If we group the work experience variable into 5 year intervals to create the discrete state variable X_t^* (i.e., $X_t^* = 1$ iff $X_t \in [0,4]$, $X_t^* = 2$ iff $X_t \in [5,9]$, etc.) and assume that wages depend only on X_t^* this does *not* reduce N^S. This is because one must know X_t to determine the value of X_{t+1}^* generated by the choice d_t. Hence, X_t has not been eliminated as a state variable. To eliminate X_t requires that one also redefine the choice variable to be whether or not to work for a five year interval. Still, discretization continues to suffer from the curse of dimensionality for the same reasons mentioned earlier, and its practical usefulness will therefore be limited to cases where the number of state variables that must be discretized is small.

Rust (1997) has proposed an ingenious simulation method that breaks the curse of dimensionality in problems with continuous state variables. The essential features of Rust's method can be illustrated using a modified version of model (3.6). Replace the law of motion $X_{t+1} = X_t + d_t$ with the absolutely continuous transition density $p(X_{t+1}|X_t, d_t)$. Assume that this density is strictly positive over the whole support of X[7]. Rust's method requires (without essential loss of generality) that the state variables

[7] These assumptions are stronger than necessary for the method to be applicable, but facilitate the example. In the context of this example, an economic interpretation of the assumptions would be that

live on the unit hypercube. So assume that $X_t \in [0,1]$, and define the uniform random variable $\eta_t \sim U[0,1]$. We will use the η_t to generate the normal wage draws via the inverse distribution function $\varepsilon_t = \sigma F^{-1}(\eta_t)$. Then Rust's algorithm would proceed as follows:

(1) Draw $\{X^m, \eta^m\}$ from the $U[0,1]^2$ distribution for $m = 1, \ldots, M$. These draws establish a grid at which the value functions will be calculated for all t.

(2) To commence the backsolving process calculate:

$$\hat{V}_T(X_T^m, \eta_T^m) = \max_{d_T} \{u_{d_T}(X_T^m, \eta_T^m)\} \qquad \text{for } m = 1, \ldots, M.$$

(3) At $t = T - 1$ calculate the alternative specific value functions:

$$\hat{V}_{T-1,d_{T-1}}(X_{T-1}^m, \eta_{T-1}^m) = u_{d_{T-1}}(X_{T-1}^m, \eta_{T-1}^m) + \delta \sum_{l=1}^{M} \hat{V}_T\left(X_T^l, \eta_T^l\right) p_M\left(X_T^l, \eta_T^l | X_{T-1}^m, \eta_{T-1}^m, d_{T-1}\right),$$

where:

$$p_M\left(X_T^l, \eta_T^l | X_{T-1}^m, \eta_{T-1}^m, d_{T-1}\right) \equiv \frac{p\left(X_T^l | X_{T-1}^m, d_{T-1}\right) p\left(\eta_T^l\right)}{\sum_{k=1}^m p\left(X_T^k | X_{T-1}^m, d_{T-1}\right) p\left(\eta_T^k\right)},$$

and set $\hat{V}_{T-1}\left(X_{T-1}^m, \eta_{T-1}^m\right) = \max_{d_{T-1}}\{\hat{V}_{T-1,d_{T-1}}\left(X_{T-1}^m, \eta_{T-1}^m\right)\}$ for $m = 1, \ldots, M$.

(4) Continue back to $t = T - 2$, and so on.

Rust refers to the $\hat{V}_t(I_t^m)$ functions as "random Bellman operators" and shows that the *expected* error in using a random Bellman operator to approximate the true Bellman operator decreases at rate $M^{1/2}$ independent of the number of state variables (S). This is not sufficient to show that the algorithm breaks the curse of dimensionality, because the expectation of the $O_p\left(M^{-1/2}\right)$ approximation errors might still increase exponentially fast with S. But Rust derives a bound on the expected error that holds uniformly (for all M, V, u, P) and that increases only linearly in S. He shows that the minimal computation cost of solving the hardest problem of dimension S within a maximum error of e has upper bound $S^4/(1-\delta)^8 e^4$. Thus, computational cost of the algorithm grows at only a polynomial rate in S. Since the rate is polynomial rather than exponential, the curse of dimensionality is removed in a formal sense (although the S^4 term implies that computational burden may still be daunting for problems with large S).

There is as yet no computational experience with this method, so we cannot comment on its performance in practice. The number of calculations required to

X_t represents human capital, and that this is continuous and evolves stochastically. From time t to $t = 1$ any change in human capital is possible, but a reasonable parameterization of $p(\cdot|\cdot,\cdot)$ would imply that increases are likely when $d_t = 1$ and decreases are likely when $d_t = 0$.

implement this procedure is proportional to $T \cdot J \cdot M^2$. The M^2 term arises because the number of draws used to simulate the future components and the number of grid points at which the future components are evaluated are set equal. This may be problematic, because in many cases it may be desirable to separately control: (1) the number of grid points, and (2) the accuracy of the future component approximation at each grid point. As Santos and Vigo-Aguiar (1998) note "... it would not be optimal to operate with a very fine grid of points in cases where the approximation errors from maximization and integration [involved in calculating the future component] are large." And in a set of numerical experiments on problems with a large but finite number of state points, Keane and Wolpin (1994) found that the most cost-effective method of achieving accurate approximate solutions was to use a rather small number of draws to simulate the Emax functions at each state point, but to include a large number of state points in the grid of points at which the Emax functions are simulated.

Rust's method is designed for problems with discrete decision variables and continuous state variables. It is not applicable to problems with discrete or mixed discrete/continuous state variables. For instance, it could not be applied to the original version of problem (3.6), in which the state variable X_t evolves deterministically according to $X_{t+1} = X_t + d_t$. Then, $p(X_{t+1}|X_t, d_t)$ is degenerate, and the interpolation method implicit in step (3) of the algorithm breaks down[8]. It is then necessary to employ an algorithm in which a value for $\hat{V}_{t+1}(X_{t+1})$ is available for any value of X_{t+1} that might be attained given $\left\{\{X_t^m, j\}_{j=1}^J\right\}_{m=1}^M$.

Keane and Wolpin (1994) present an algorithm that can be applied to discrete dynamic optimization problems with either discrete or continuous state variables. In the discrete case the algorithm involves: (1) using Monte Carlo integration to simulate the Emax functions at only a subset of the total number N^S of relevant $\{I_t^*, d_t\}$ combinations, and (2) interpolation of the non-simulated Emax values using a regression function. To describe this method more precisely it is necessary to first establish some notation. Index the set of unique I_t^* values by $s = 1, \ldots, N_t^S$ and denote the elements of this set by I_{ts}^*. Choose the number $G < N_T^S$ of state points at which the F_t functions will be simulated. Note that in many examples N_t^S is increasing with t, and for some t^* it is the case that $G \geqslant N_{t^*}^S$. The backsolving process proceeds as follows.

(1a) Draw G integers from the multinomial distribution with equal probability on $\left\{1, \ldots, N_T^S\right\}$. Denote the chosen integers by $\{s(g)\}$, $g = 1, \ldots, G$.

(1b) Form the simulated Emax functions:

$$\hat{F}_T\left(I_{T,s(g)}^*\right) = \delta \hat{E}\left[\max_{d_T}\left\{u_{d_T}(I_T)\right\} |I_{T,s(g)}^*\right] \quad \text{for } g = 1, \ldots, G.$$

[8] Furthermore, even if X_t is continuous, if $p\left(X_{t+1}|X_t, d_t\right)$ is not strictly positive over a substantial part of the range of X then $p_M\left(X_t^l|X_{t-1}^m, d_{t-1}\right)$ will often be zero and, as a practical matter, the interpolation in step 3 will again break down, unless M is very large.

(Note: possible methods of forming the simulators $\hat{E}$ were discussed earlier.)

(1c) Run a regression of the $\hat{F}_T\left(I^*_{T,s(g)}\right)$ values on functions of the arguments $I^*_{T,s(g)}$.

(2a) Draw G integers from the multinomial distribution with equal probability on $\left\{1, \ldots, N^S_{T-1}\right\}$. Again denote the chosen integers by $\{s(g)\}$, $g = 1, \ldots, G$.

(2b) Form the simulated Emax functions:

$$\hat{F}_{T-1}\left(I^*_{T-1,s(g)}\right) = \delta\hat{E}\left[\max_{d_{T-1}}\left\{u_{d_{T-1}}\left(I_{T-1,s(g)}\right) + \hat{F}_T\left(I^*_{T-1,s(g)}, d_{T-1}\right)\right\} | I^*_{T-1,s(g)}\right].$$

If $\hat{F}_T\left(I^*_{T-1,s(g)}, d_{T-1}\right)$ was calculated in step (1b); i.e., if $\left(I^*_{T-1,s(g)}, d_{T-1}\right) \in \left\{I^*_{T,s(g)}\right\}_{g=1}^{G}$; then use that value. If not, then use the regression function fit in step (1c) to interpolate the needed value.

(2c) Run a regression of the $\hat{F}_{T-1}\left(I^*_{T-1,s(g)}\right)$ values on functions of the arguments $I^*_{T-1,s(g)}$.

(3) Continue back to $T-2$, and so on. If a t^* is reached such that $G \geqslant N^S_{t^*}$ then the Emax functions are simulated at all state points and steps (a) and (c) are not necessary.

In the case of continuous state variables step (a) of the algorithm must be modified. Draw the continuous state variables from a density that has positive mass over the whole feasible domain of those state variables (e.g., a uniform distribution could be used as in Rust's algorithm). Also, in the (b) steps interpolation will *always* be necessary, provided there are no atoms in the transition densities $p\left(I^*_t | I^*_{t-1}, d_{t-1}\right)$.

Keane and Wolpin (1994) compare the performance of alternative interpolation functions for use in steps (1c) and (2c). They present numerical examples where $N^S_T = 13\,150$ and $N^S \approx 130\,000$, and where a very precise solution to the dynamic optimization problem required approximately 50 minutes on a Cray-2. Approximate solutions using their algorithm with $G = 500$, and with crude frequency simulation of the Emax functions (as in Equation 3.5) with $M = 2000$, required only 6 cpu seconds – a 500 fold speed improvement. When the resulting approximate decision rules were simulated, they generated choice behavior that was in close agreement with behavior based on the "true" decision rules (i.e. 96% to 99% choice agreement in the examples considered). Furthermore, wealth losses from using the approximate rather than "true" decision rules were on the order of a few tenths of a percent of lifetime wealth for the simulated agents. When the approximate solution algorithm was embedded in a maximum likelihood estimation algorithm, and parameter estimates obtained using data simulated using the "true" decision rules, those estimates were in almost all cases quite close to the true parameter values. Successful empirical applications of the Keane–Wolpin algorithm include Erdem and Keane (1996) and Keane and Wolpin (1997, 2000a,b)[9].

Rust (1997) notes that the Keane–Wolpin algorithm does not break the curse of dimensionality in the continuous state variable case because the (c) steps involve fitting an approximation to a multivariate function, and continuous multivariate function approximation is subject to a curse of dimensionality [see Traub, Wasilkowski and Wozniakowski (1988)]. That is, as the number of state variables grows large the minimum computation time needed to approximate the Emax functions within a given tolerance may grow exponentially – under a worst case scenario[10]. It is important to stress however, that this worst case analysis applies to very general classes of functions[11]. In fitting approximations to the Emax function in any particular application we are not dealing with an arbitrary unknown function, but rather with a function about which a great deal is known [see e.g., Stern (1991)][12]. Hence, additional research is clearly needed on the properties of polynomial (and other) approximations to Emax functions.

Another point to note is that the fitting of polynomial or other approximations to the Emax functions requires a trivial amount of computation time relative to the calculation of the Emax functions themselves. This was the original motivation behind the Bellman, Kalaba and Kotkin (1963) and Keane and Wolpin (1994) approaches. In the case of polynomial approximation, if the number of terms in the polynomial is increased, the significant increase in computation cost does *not* arise from additional time needed to fit the polynomials. Rather, it arises because, in order to obtain sufficient

[9] In the Bellman et al. (1963) approach, rather than choosing a random set of points at which to evaluate the F_t functions, the state variables are assumed to lie in the $[0, 1]$ interval and the Gaussian quadrature points based on Legendre polynomials are used. The $\hat{F}_t$ values calculated at the quadrature points are then regressed on the polynomial terms in the state variables. Their algorithm shares with the Keane–Wolpin algorithm the essential feature that the fitted values $\hat{F}_t$ are then substituted for the future component at each successive step of the backsolving process. A key difference is that, if one uses quadrature points rather than randomly chosen grid points, the number of grid points is R^S, where R is the number of quadrature points, and hence grows exponentially with S.

[10] For instance, consider polynomial approximations. With S state variables a fully interacted polynomial of order k has $1 + Sk + k^2S(S - 1)/2$ terms. Thus, even if one must set $k \propto S$ to maintain accuracy within a given tolerance as S increases, computation time grows only at the polynomial rate S^4, not exponentially. Hence, the Traub et al. (1988) results suggest that a faster rate of growth of K may be necessary in the worst case.

[11] For example, Traub et al. (1988) show that for the class of nonperiodic L_2 functions $f : [0, 1]^d \to \Re$ which are r times continuously differentiable, the worst case computation time to achieve an ε approximation is proportional to $\varepsilon^{-1/r} \left(\ln\varepsilon^{-1}\right)^{d-1}$.

[12] A common special case is when the current period payoffs have only additive errors, so $u_j(I_t) = u_j(\bar{I}_t) + \varepsilon_{tj}$. Then we can write $V_j(I_t) = V_j(\bar{I}_t) + \varepsilon_{tj}$ and the Emax function evaluated at $\bar{I}_t$ depends only on the J arguments $V_j(\bar{I}_t)$ for $j = 1, \dots, J$, and not on values of the underlying state variables $\bar{I}_t$. Hence, the dimensional of the multivariate function to be approximated remains J regardless of the number of state variables S.

data to fit a larger polynomial, it is necessary to evaluate the Emax functions at a larger sample of state points.

For purposes of econometric work, the approximate solution of dynamic optimization models is not an end in itself. Rather, we wish to estimate the parameters of such models. In some contexts the parameters themselves may be the objects of interest, and in others the goal may be to use the estimated model to predict or simulate behavior. The approximate solution of the dynamic optimization problem is merely an input into the estimation process. Since dynamic optimization models are typically highly nonlinear, this process will typically involve iterative search – that is, it will require that the optimization problem be solved, and the resultant econometric objective function be evaluated, at many trial parameter values.

A radically different approach to econometric work on discrete dynamic optimization models is to seek ways to circumvent the solution of the optimization problem entirely, while still learning about parameters of interest and/or the agent's decision rules. Algorithms that cut the Gordian knot in this way were developed in the pioneering work by Hotz and Miller (1993) and Manski (1991). The basic idea of these algorithms is to use observed outcomes to *estimate* the values of the $F_{t+1}(I_t^*, j)$ for different state/choice combinations, rather than *solving* agents' dynamic optimization problem to obtain the $F_{t+1}(I_t^*, j)$. Given such estimates, one can identify the structural parameters of the current payoff functions.

There are, however, two important limitations of these approaches. First, since actual outcomes are used to estimate agents' expectations at the time decisions were made, the methods require the assumption of a stationary environment[13]. Second, since outcomes for *all* agents in a given state are used to estimate the expectations of each agent in that state, the methods cannot accommodate unobserved state variables[14]. The stationarity assumption combined with the no unobserved state variables assumption means that models estimated using these algorithms cannot be used to predict the impact of policy interventions or regime changes that are not already present in the data and captured by observed state variables.

Geweke and Keane (1995) proposed an alternative approach to inference in discrete dynamic optimization models that also circumvents the need to solve the dynamic optimization problem. This approach involves polynomial approximation of the future component in Equation (3.2), and it avoids the stationarity and no unobserved state variables assumptions of the Hotz–Miller and Manski approaches. Geweke and Keane consider situations in which the econometrician is willing to assume a parametric functional form for the current period payoff functions $\left(u_{d_t}(I_t)\right)$ and the law of motion of the state variables (M). These type of assumptions are also required in all the other

[13] Not only must expectations be rational, they must also be fulfilled – a much stronger assumption. See Manski (1991) and also Keane and Runkle (1990) for further discussion.

[14] An exception is the very special case where the unobserved state variable affects only current and not future payoffs.

methods we have considered so far. The difference in Geweke and Keane (1995) is in how the future component is interpreted. In the equation for the future component:

$$F_{t+1}(I_t, d_t) = E\left[\sum_{\tau=t+1}^{T} \delta^{\tau-t} u_{d_\tau}(I_\tau) \,|I_t, d_t\right], \tag{3.7}$$

all the methods we have reviewed so far assume that $E[\cdot|\cdot]$ is the mathematical expectation operator. In the present context this is equivalent to assuming that agents form expectations optimally given the structure of the model (i.e., that they have rational expectations). In contrast, Geweke and Keane consider a context in which the econometrician does not make strong assumptions about how subjective expectations are formed. Rather he/she specifies $F_{t+1}(I_t, d_t) = F_{t+1}\left(I^*_{t+1}\right)$ as a flexible polynomial function of the relevant state variables I^*_{t+1}. Denote this function by $F_{t+1}\left(I^*_{t+1}(I_t, j)\,|\pi\right)$. We then have:

$$V_j(I_t) = u_j(I_t|\theta) + F_{t+1}\left(I^*_{t+1}(I_t, j)\,|\pi\right). \tag{3.8}$$

Here, θ is the vector of structural parameters of the current period payoff function, and π is a vector of polynomial coefficients that characterize expectation formation.

The polynomial approximation methods described by Bellman, Kalaba and Kotkin (1963), Keane and Wolpin (1994), Marcet (1994), Judd (1992) and others, all assume that $E[\cdot|\cdot]$ in Equation (3.7) is the mathematical expectation operator. This means that the true $F_{t+1}(I_t, d_t)$ is determined by θ. These methods seek a π vector that gives approximations $F_{t+1}\left(I^*_{t+1}(I_t, j)\,|\pi\right)$ that are in some sense close to the projection of the F_{t+1} on the space spanned by polynomials of given order.

In contrast, the fact that Geweke and Keane do not assume that $E[\cdot|\cdot]$ in Equation (3.7) is the mathematical expectation operator has an important consequence: π becomes a free parameter that can be estimated from data, provided θ and π are jointly identified. Given data on choices, along with at least some information on current payoffs, it is often possible to identify both θ and π, and thus to learn both about the payoff function parameters and the structure of expectations from the data[15].

A leading example where identification of both θ and π will be achieved is when the payoff is observed if and only if an alternative is chosen. In this case, after substitution of a flexible polynomial function for the future component as in Equation (3.8),

[15] Of course, any combination (θ, π) can be rationalized as optimal behavior by choosing an appropriate structure of payoffs at $T+1$ and extending the summation in Equation (3.7) out to $T+1$, provided the $T+1$ payoffs are allowed to be functions even of "irrelevant" state variables (e.g., the whole history of choices $\{d_1, \ldots, d_T\}$ in model (3.6), rather than only X_{T+1}). One interpretation of our procedure is that we continue to assume that $E[\cdot|\cdot]$ in Equation (3.7) is the mathematical expectation operator, but that we leave the payoff functions at $T+1$ unspecified.

the dynamic discrete choice model takes on a form similar to a static Roy (1951) model augmented to include influences on choice other than the current payoffs, as in Heckman and Sedlacek (1985). The difference is that Equation (3.8) incorporates restrictions on the form of the non-payoff components $F_{t+1}\left(I^*_{t+1}(I_t,j)\,|\pi\right)$ that are implied by the dynamic optimization model and that are not typically invoked in the estimation of static selection models. First, the model implies that the parameters π of the non-payoff component of the value function are constant across alternatives. Second, the model also implies that the regressors $I^*_{t+1}(I_t,j)$ that enter the non-payoff component vary in a systematic way across alternatives that is determined by the law of motion for the state variables. Typically, these two features of the model are manifested in cross-equation restrictions on θ and π that are not operative in static selection models.

This point is illustrated in a simple example. Let $J = 3$ and assume payoffs are given by $W_{tj} = X_{t1}\beta_{1j} + X_{t2}\beta_{2j} + \varepsilon_{tj}$ for $j = 1,2$ where $\varepsilon_t \sim N(0,\Sigma)$, and $W_{t3} = 0$. Further assume that the laws of motion for the scalar state variables X_{tj} are $X_{t+1,j} = X_{t,j} + \chi\,[d_t = j]$ for $j = 1,2$, where $\chi[\cdot]$ is an indicator function. Letting the future component $F_{t+1}\left(X_{t+1,1}, X_{t+1,2}\right)$ be a second order polynomial in the two state variables, we obtain the alternative specific value functions:

$$\begin{aligned}
V_{it1} &= X_{it1}\beta_{11} + X_{it2}\beta_{21} + \varepsilon_{it1} + (X_{it1}+1)\,\pi_1 + (X_{it1}+1)^2\,\pi_2 + X_{it2}\pi_3 + X^2_{it2}\pi_4 \\
&\quad + (X_{it1}+1)\,X_{it2}\pi_5, \\
V_{it2} &= X_{it1}\beta_{12} + X_{it2}\beta_{22} + \varepsilon_{it2} + X_{it1}\pi_1 + X^2_{it1}\pi_2 + (X_{it2}+1)\,\pi_3 + (X_{it2}+1)^2\,\pi_4 \\
&\quad + X_{it1}(X_{it2}+1)\,\pi_5, \\
V_{it3} &= X_{it1}\pi_1 + X^2_{it1}\pi_2 + X_{it2}\pi_3 + X^2_{it2}\pi_4 + X_{it1}X_{it2}\pi_5,
\end{aligned}$$

where $i = 1, \ldots, I$ indexes agents. The decision rule is $d_{it} = j$ iff $V_{ij} - V_{ik} \geqslant 0$ for all $k \neq j$. Joint identification of $\theta = (\beta, \Sigma)$ and π in this case is obvious. Observe that the latent indices in the reduced form probit selection rule are:

$$Z_{itj} \equiv V_{itj} - V_{it3} = \phi_{1j} + X_{it1}\phi_{2j} + X_{it2}\phi_{3j} + \varepsilon_{itj} \quad \text{for} \quad j = 1,2, \tag{3.9}$$

and $Z_{it3} = 0$, where $\phi_{11} \equiv \pi_1 + \pi_2$, $\phi_{21} \equiv \beta_{11} + 2\pi_2$, $\phi_{31} \equiv \beta_{21} + \pi_5$, $\phi_{12} \equiv \pi_3 + \pi_4$, $\phi_{22} \equiv \beta_{12} + \pi_5$ and $\phi_{32} \equiv \beta_{22} + 2\pi_4$. The parameters β_{11}, β_{21}, β_{12}, β_{22} and Σ can be consistently estimated via joint estimation of the selection rule (3.9) with the equations for $\{W_{itj}\}^2_{j=1}$, relying for identification on joint normality of ε_{itj} for $j = 1,2$, and on the cross-equation restriction $\phi_{22} - \phi_{31} = \beta_{12} - \beta_{21}$. The later is an example of the type of restriction in the dynamic model that would not arise in a static selection model[16].

[16] As a practical matter, such cross-equation restrictions are critical for identification in the model, despite the fact that they are not necessary for formal identification. In Section 6, we present Monte-Carlo results for static selection models which show how, in the absence of parameter restrictions, the

Next, given the estimates of the payoff function parameters, it is obvious that π_1 through π_5 are identified from estimation of the "structural" probit model that is obtained from Equation (3.9) after substituting in the estimates of β_{11}, β_{21}, β_{12} and β_{22} from stage 1. Observe that the latent indices in the structural probit are:

$$\begin{aligned} Z_{it1}^* &= X_{it1}\,\hat{\beta}_{11} + X_{it2}\,\hat{\beta}_{21} + \varepsilon_{it1}^* + \pi_1 + (2X_{it1} + 1)\ \pi_2 + X_{it2}\pi_5 \\ Z_{it2}^* &= X_{it1}\,\hat{\beta}_{12} + X_{it2}\,\hat{\beta}_{22} + \varepsilon_{it2}^* + \pi_3 + (2X_{it2} + 1)\ \pi_4 + X_{it1}\pi_5. \end{aligned}$$

Thus, the model implies the restriction that the coefficient on X_{it2} in the first equation and the coefficient on X_{it1} in the second equation are equal. Note that use of higher order polynomial approximations to the future component would generate additional cross-equation restrictions of this type.

The parameters θ and π remain jointly identified if the payoffs are given by $W_{itj} + \eta_{itj}$ where η_{itj} is a component of the payoff that is never observed by the econometrician, provided that η_{itj} is distributed independently of the $\{W_{itj}\}_{j=1}^{2}$, for this generates a switching regression model like that in Lee (1978, 1979).

In Geweke and Keane (1995) and Geweke, Houser and Keane (1998) we present a series of Monte-Carlo experiments in which this approach of approximating the future component by a polynomial is applied to data generated from various discrete dynamic optimization models. Once the future component is specified as a polynomial, statistical inference via maximum likelihood or use of Bayesian methods is no more difficult than in static selection or switching regression models. In our experiments, we find that this approach leads to reliable inferences about the parameters θ that enter the current payoff functions, and about the parameters π that characterize expectations. In particular, we find that if the true future components are in fact generated by a "correct" solution of the dynamic optimization problem (i.e., if agents use the optimal decision rule) the use of reasonably low order polynomial approximation still results in reliable inferences about θ. And the resulting approximations to the optimal decision rules typically generate present value of lifetime payoff losses on the order of a tenth of one percent when used in place of the optimal rules. We present some representative results from these experiments in Section 7.

There are several appealing aspects of the polynomial approximation approach we have described here. First, for discrete state variables that take on only a finite configuration of values, a finite order polynomial can generate any (finite valued) future component exactly. For continuous state variables, the Stone–Weierstrass Theorem states that if the future component is a continuous function of the state variables whose

likelihood contains little information about either payoff function parameters or the correlation between payoff and choice equation errors. However, the use of parameter restrictions allows all these parameters to be pinned down precisely. A nice feature of the approach to estimation of dynamic selection models that we describe here is that cross-equation restrictions that aid in identification arise naturally from the structure of the model.

domain is a compact subset of $\Re^2$ then a uniform approximation of arbitrary accuracy can be achieved by use of a sufficiently high order polynomial.

Second, the use of finite order polynomial approximation means that in general Equation (3.8) will be a misspecification of agents' decision rules. But it is well known that in this case the pseudo-MLE converges almost surely to the subset of the parameter space on which the Kullback–Leibler distance (KL) between the true data distribution and the distribution generated by the approximate decision rule is minimized [see Huber (1967), Pfanzagl (1969)]. Recently Bunke and Milhaud (1998) proved an analogous result for pseudo-Bayes estimators with respect to general loss functions. They show that in the case of a unique pseudo-true parameter vector that minimizes KL, pseudo-Bayes estimators are strongly consistent.

Third, unlike the Hotz and Miller and Manski approaches, this method does not require stationarity assumptions. This is because agents' expectations of future payoffs are inferred from *current* choices and payoffs, rather than future payoff realizations. Fourth, the method can accommodate unobserved heterogeneity, since parametric heterogeneity distributions are generally identified in static selection models. Fifth, the method does not require that each payoff function have a unique additive error, and it can accommodate flexible specification of error distributions using methods like those we implement in Section 6 for the static selection model. Sixth, unlike all the methods we have described so far, this method can be extended to handle joint discrete/continuous decision variables, as demonstrated in Houser (1998).

Finally, the method allows more flexibility in analysis of regime shifts than do the Hotz and Miller and Manski approaches. Behavior of agents under regimes not present in the observed data can be simulated, provided the regime shift can be represented as a change in observed state variables in agents' decision rules. That is, the data must contain some variation along the dimensions of interest, but need not contain the exact configuration of the state variables that characterize the new regime.

4. Classical simulation estimation of the multinomial probit model

The early work on simulation estimation in econometrics was concerned primarily with the multinomial probit model (MNP). So we begin with a description of MNP. Let j index mutually exclusive alternatives from the set $\{1, \ldots, J\}$. The utility that agent i receives from choice of alternative j is specified as:

$$y_{ij}^* = W_{ij}\Gamma_j + \varepsilon_{ij}^* \qquad j = 1, \ldots, J, \quad i = 1, \ldots, N,$$

where W_{ij} is a vector of covariates, Γ_j is a conformable vector of coefficients, and $\varepsilon_i^* \equiv \left(\varepsilon_{i1}^*, \ldots, \varepsilon_{iJ}^*\right)' \sim N(0, \Sigma^*)$. The econometrician does not observe the utilities $\{y_{ij}^*\}_{j=1}^J$ but only the covariates and the decision d_i where $d_i = j$ iff $y_{ij}^* \geqslant y_{ik}^* \; \forall \; k$.

It is useful to distinguish between two types of covariates: those that vary across alternatives for an individual, and those that vary only across agents. We denote the former by Z_{ij} and latter by X_i and rewrite the model:

$$y_{ij}^* = Z_{ij}\,\gamma_j + X_i\,\beta_j^* + \varepsilon_{ij}^* \qquad j = 1, \ldots, J, \quad i = 1, \ldots, N, \tag{4.1}$$

where γ_j and β_j^* are obtained from Γ_j in the obvious way. Numerical experiments in Keane (1992) indicate that identification of MNP parameters is extremely tenuous in the absence of covariates like the Z_{ij}, so we include such covariates in all the MNP examples that we discuss, both here and in Sections 6 and 7.

The parameters of Equation (4.1) are not identified, because only utility differences affect choices, and choices are invariant to the scale of utilities. A common normalization is to define J as a base alternative, and define:

$$\begin{aligned} y_{ij} &\equiv y_{ij}^* - y_{iJ}^* = Z_{ij}\,\gamma_j - Z_{iJ}\,\gamma_J + X_i\left(\beta_j^* - \beta_J^*\right) + \left(\varepsilon_{ij}^* - \varepsilon_{iJ}^*\right) \\ &\qquad\quad = Z_{ij}\,\gamma_j - Z_{iJ}\,\gamma_J + X_i\,\beta_j + \varepsilon_{ij} \qquad j = 1, \ldots, J-1, \\ y_{iJ} &= 0, \end{aligned}$$

where $\varepsilon_i \equiv (\varepsilon_{i1}, \ldots, \varepsilon_{iJ-1}) \sim N(0, \Sigma)$ and Σ is a $(J-1) \times (J-1)$ covariance matrix obtained from Σ^*. Further, the scale normalization is usually imposed by setting $\Sigma_{11} = 1$.

It will be convenient to write $y_{ij} = \bar{y}_{ij} + \varepsilon_{ij}$ for $j = 1, \ldots, J$ and adopt the convention that $\bar{y}_{iJ} = \varepsilon_{iJ} = 0$. Then, agent i chooses option j iff $y_{ik} - y_{ij} \leqslant 0\ \forall k$, which generates the $J-1$ dimensional partition of the ε space $\varepsilon_{ik} - \varepsilon_{ij} \leqslant \bar{y}_{ij} - \bar{y}_{ik}\ \forall k$. Define $\tilde{\varepsilon}_{ik}^j = \varepsilon_{ik} - \varepsilon_{ij}$ for $k = 1, \ldots, J$, and further define $\tilde{\varepsilon}_i^j = \left(\tilde{\varepsilon}_{i1}^j, \ldots, \tilde{\varepsilon}_{i,j-1}^j, \tilde{\varepsilon}_{i,j+1}^j, \ldots, \tilde{\varepsilon}_{iJ}^J\right) \sim N(0, \tilde{\Sigma}^j)$. Then, the probability that agent i chooses option j can be written as the $J-1$ dimensional integral:

$$\begin{aligned} p(j|Z_i, X_i, \gamma, \beta, \Sigma) &= \int_{-\infty}^{\bar{y}_{ij} - \bar{y}_{i1}} \cdots \int_{-\infty}^{\bar{y}_{ij} - \bar{y}_{iJ}} p\left(\tilde{\varepsilon}_1^j, \ldots, \tilde{\varepsilon}_J^j | \tilde{\Sigma}^j\right) \mathrm{d}\tilde{\varepsilon}_J^j \cdots \mathrm{d}\tilde{\varepsilon}_1^j \\ &= P\left(\bar{y}_{ij} - \bar{y}_{i1}, \ldots, \bar{y}_{ij} - \bar{y}_{iJ} | \tilde{\Sigma}^j\right). \end{aligned} \tag{4.2}$$

Letting $d = (d_1, \ldots, d_J)$ the likelihood function is then:

$$p(d|\,\gamma, \beta, \Sigma) = \prod_{i=1}^{N} \prod_{j=1}^{J} P\left(\bar{y}_{ij} - \bar{y}_{i1}, \ldots, \bar{y}_{ij} - \bar{y}_{iJ} | \tilde{\Sigma}^j\right)^{\chi[d_j = j]}. \tag{4.3}$$

Thus, the key computational problem in MNP estimation is that construction of the likelihood requires evaluation of $J-1$ dimensional integrals. And, unfortunately, deterministic methods for evaluation of integrals (such as quadrature) suffer from a curse of dimensionality, in that computation time required to insure a given level of accuracy grows exponentially with dimension. Given current computer speeds, estimation of MNP models using deterministic methods is only feasible for $J = 3$ or perhaps 4.

This statement may seem surprising, because it is now certainly feasible to evaluate quite high dimensional integrals to high accuracy within reasonable time using quadrature. The key point to note, however, is that maximum likelihood estimation of the MNP requires that the likelihood be evaluated at *many* trial values for the $K \times 1$ parameter vector $\theta \equiv (\gamma, \beta, \Sigma)$. At each trial value θ_T, the $J - 1$ dimensional integrals in Equation (4.2) must be evaluated for all N agents in the population. Furthermore, note that on each iteration of a derivative based search algorithm designed to locate $\hat{\theta}_{ML}$, it is necessary to evaluate the likelihood at several different values of the parameter vector. These are 1) the initial trial parameter vector θ_T, 2) K bumped parameter values $\theta_T + b\Delta_k$ (where Δ_k for $k = 1, \ldots, K$ is a vector with a one in the kth position and zeros elsewhere, and b is a bump size) in order to construct numerical derivatives of the likelihood, and 3) the new trial parameter vectors θ_T', of which a line search algorithm will always try out at least two. So each iteration will involve at least $K + 3$ evaluations of the likelihood.

To give an idea of the computational burden involved, consider a rather small sized problem in which $J = 4$ and $N = 500$. With scalar Z_{ij} and X_i there are 12 parameters ($\{\gamma_j\}_{j=1}^4$, $\{\beta_j\}_{j=1}^3$ and the 5 free elements of Σ). If 50 iterations are required for convergence, then approximately $50 \cdot (12 + 3) \cdot 500 = 375\,000$ three-dimensional integrals must be evaluated. Even if a computer could evaluate 100 such integrals per cpu second, total computation time would be over an hour[17]. The burden increases rapidly as sample size, number of covariates and number of iterations increase[18].

Recognition of this problem led a number of investigators to consider simulation-based estimation for the MNP. One approach is to simulate the integrals in Equation (4.2) using fast Monte-Carlo methods, and to insert these approximations into the likelihood (4.3). This gives the simulated maximum likelihood (SML) estimator.

[17] Such a timing figure is in fact only realistic (given currently available computers) if simulation methods are employed to evaluate the integrals. For instance, the Fortran code for the GHK simulator discussed in Section 2, requires approximately .01 cpu seconds to evaluate a three dimensional integral using 100 draws, on a Sparc Ultra-2 workstation. Thus, approximately 100 integrals could be evaluated per cpu second. If quadrature methods were used instead, far greater computation time would be necessary to achieve reasonable accuracy.

[18] A general expression for the approximate number of integrations required is $C \cdot N \cdot \{J \cdot N_z + (J - 1) \cdot N_x + J(J - 1)/2 + 2\}$ where C is the number of iterations, $N_z = \dim \{Z_{ij}\}$, and $N_x = \dim \{X_i\}$. Notice that the number of parameters increases rapidly in J, due to the J^2 term arising through the number of covariance parameters. This has the side effect of increasing the number of iterations C required for convergence, and the sample size N required to obtain reliable estimates. There have been proposals to impose low dimensional factor structures on the covariance matrix to avoid this proliferation of parameters [see Elrod and Keane (1995), Geweke, Keane and Runkle (1994)].

Early work on SML was reported by Albright, Lerman and Manski (1977) and Lerman and Manski (1981). They used crude frequency simulators of the form:

$$\hat{p}_F(j|Z_i,X_i,\theta) = M^{-1}\sum_{m=1}^{M}\chi\left[\tilde{\varepsilon}_1^{j(m)} \leqslant \bar{y}_{ij}-\bar{y}_{i1},\ \ldots,\ \tilde{\varepsilon}_J^{j(m)} \leqslant \bar{y}_{ij}-\bar{y}_{iJ}\right],$$

where $\{\tilde{\varepsilon}_k^{j(m)}\}_{k=1}^{J}$ for $m = 1,\ldots,M$ are i.i.d. draws from the joint distribution $\tilde{\varepsilon}_i^j \sim N(0,\tilde{\Sigma}^j)$. The general consensus regarding this work is that crude frequency simulation did not perform satisfactorily. The two main problems are: 1) that low probability events are often simulated to have zero probability, even for reasonably large M, which sends the likelihood to zero, and 2) that frequency simulators are not differentiable (or even continuous) functions of θ. Hence, the use of frequency simulation renders the likelihood a step function with jumps at parameter configurations that produce ties among alternatives. Pakes and Pollard (1989) present consistency and asymptotic normality results that can be applied in cases where the objective function is discontinuous. However, as a practical matter, if the objective function is a step function it precludes use of gradient based search algorithms and forces the econometrician to resort to non-gradient methods (like simplex) that are typically much slower. The most notable empirical application of SML using frequency simulation is by Pakes (1986), who overcame these problems by using very large simulation sizes.

A further concern with SML is that the simulation errors for the individual choice probabilities (4.2) enter the likelihood function (4.3) nonlinearly. Hence, if simulation size M is held fixed as sample size N is increased, parameter estimates are inconsistent. It is necessary that $M \to \infty$ as $N \to \infty$ in order for SML to be consistent [see Pakes and Pollard (1989)]. But the real issue is that, for fixed N, one must choose M large enough to render the small sample bias negligible (or tolerable). Crude frequency simulators are quite imprecise, so very large M may be necessary to achieve this goal.

A major breakthrough in the simulation estimation literature occurred when McFadden (1989) proposed the use of probability simulators that are smooth functions of the model parameters. Smooth simulators have many critical advantages over crude frequency simulators. They can be constructed so as to be both unbiased and bounded away from zero for any M, and they deliver a simulated likelihood that is a differentiable function of the model parameters – thus allowing use of gradient based search algorithms. Furthermore, smooth simulators can be far more efficient than crude frequency simulators in terms of the accuracy that is achieved for given computation time. This opened the possibility that for MNP with large J it would be feasible to construct SML estimators with tolerably small simulation induced bias.

Lee (1992, 1995) has examined the asymptotic properties of the SML estimator for discrete choice models when smooth simulators are employed. Since the simulated likelihood function is differentiable in this case, traditional asymptotic methods can be used, as opposed to the more sophisticated empirical process methods used in Pakes and Pollard (1989). Lee shows that consistency of SML only requires that $M \to \infty$ as $N \to \infty$, with no particular rate required on M. But it is necessary to have $M/\sqrt{N} \to \infty$ as $N \to \infty$ in order for $\hat{\theta}_{\text{SML}}$ to have an asymptotically normal limiting distribution that is properly centered at zero (this condition also guarantees asymptotic efficiency). If M increases at a slower rate than $\sqrt{N}$, then $\sqrt{N}(\hat{\theta}_{\text{SML}} - \theta)$ diverges and $\sqrt{N}$ consistency is not achieved[19]. Further, Lee (1995) points out that the leading bias term in the asymptotic expansion of the MLE is $O_p\left(N^{-1/2}\right)$, while the SML estimator has an additional bias term that is $O_p\left(N^{1/2}M^{-1}\right)$. Hence, if M grows at a slower rate than $\sqrt{N}$, the bias of SML is more severe than for the MLE. For given N and M, the expected value of this bias term is a function of the variance of the probability simulator that is used. Clearly then, the use of a smooth simulator that achieves low variance with reasonable computational cost is essential for SML to be operational.

There are many varieties of smoothed unbiased probability simulator, of which McFadden (1989) presents several. Choice of simulator is crucial, because it is easy to construct realistic examples in which well-known methods produce dismal results [see Geweke (1989)]. The current consensus of the literature is that the GHK recursive simulator, which we described in Section 2, is the most reliable and accurate general purpose smooth simulator among those methods that are currently available. This method was developed both by Keane (1990, 1993, 1994) and in an independent line of research due to Geweke (1991) and Borsch-Supan and Hajivassiliou (1993)[20]. Evaluations of the relative performance of GHK versus alternative simulators in terms of root mean square error (RMSE) per given computation time include Hajivassiliou, McFadden and Ruud (1996), Vijverberg (1997) and Andrews (1999). These studies consider experimental designs in which the number of alternatives in the choice set, covariance structure of the errors, and size of the probability to be simulated are varied. Ranking methods by the RMSE criterion, GHK is usually first regardless of the treatment. In cases where GHK is not first it is outperformed only marginally by other methods. Furthermore, those methods that can marginally outperform GHK under certain treatments are typically found to perform quite poorly under other treatments. Andrews (1999) shows that use of antithetic acceleration often leads to substantial improvements in the performance of GHK. To implement antithetic acceleration,

[19] In fact, the rate that is achieved is exactly M. Lee (1995) also shows that if $M/\sqrt{N} \to \lambda$ as $N \to \infty$, where λ is a positive constant, then $\sqrt{N}(\theta_{\text{SML}} - \theta) \xrightarrow{d} N(\lambda \Omega \bar{\mu}, \Omega)$ where Ω is the inverse of the information matrix and $\bar{\mu}$ is a term related to the simulation error variance. Thus, SML is consistent but suffers from an asymptotic bias when M grows at exactly a $\sqrt{N}$ rate. We would like to point out that the Geweke, Keane and Runkle (1994, 1997) papers contain the misleading statement that "SML is consistent if $M/\sqrt{N} \to \infty$ as $N \to \infty$". This is true, but as the above discussion makes clear, slower rates suffice for consistency alone. We should have said that SML is $\sqrt{N}$ consistent and asymptotically normal (with a limiting distribution properly centered at zero) if and only if $M/\sqrt{N} \to \infty$ as $N \to \infty$.

[20] The acronym GHK (Geweke, Hajivassiliou, Keane) was coined by McFadden at the 1990 Invitational Choice Symposium in Banff, Alberta.

simply take the average of two GHK simulators, based on the uniform random variables u and $\iota - u$, where ι is a conformable vector of ones.

Despite these encouraging results, our main concern is with the small sample properties of SML when GHK is used to simulate choice probabilities. Specifically, is GHK sufficiently accurate to render the simulation induced bias in SML tolerably small? Studies of this question are Keane (1994), Geweke, Keane and Runkle (1994, 1997) and Lee (1995, 1997). The results are generally rather encouraging. For instance, Geweke et al. (1994) consider a 7 alternative MNP of the form

$$y_{ij} = \beta_{j1} + \beta_{j2}X_i + \gamma\left(Z_{ij} - Z_{i7}\right) + \varepsilon_{ij} \qquad j = 1, \ldots, 6,$$

and $y_{i7} = 0$, where the covariates and errors are designed to have a structure similar to the Nielsen scanner data on household ketchup purchases analyzed in Keane (1997). More specifically, 50 artificial data sets with $N = 5000$ each were constructed, using as covariates the household sizes and price variables from the first 5000 purchase occasions in the Nielsen data, and using as model parameters the brand intercepts, household size and price coefficients, and covariance matrix elements estimated from the Nielsen data. Rather than estimating Σ, estimates were obtained for the elements of the lower triangular Cholesky decomposition of Σ[21]. Denote these by $\left\{\{a_{jk}\}_{k=1}^{j}\right\}_{j=2}^{6}$ and note that the scale normalization was imposed by setting $a_{11} = 1$. Choice probabilities were simulated using $M = 30$ draws. This is a very difficult example (perhaps the most stringent test to which SML has been subjected in the literature) for three reasons: 1) inspection of the $\{a_{jk}\}$ reveals that the variances are very unequal across alternatives, a treatment that causes difficulty for all simulators, 2) conditional choice probabilities for some alternatives (especially 2, 4 and 5) are often quite small, and 3) the covariates are rather ill-behaved, in that prices of all brands tend to move together. The results are presented in Table 4.1, which is generated from Geweke et al. (1994, Table 10).

Note that the t-statistic for the estimated bias is significant for 22 out of 33 model parameters. Given a large enough number of replications, even a quantitatively small bias will be significant. One way to gauge the practical importance of the biases is to look at the ratio of the bias in a parameter to the typical value of its asymptotic standard error estimate (ASE). This is reported in the last column of the table. For only two of the brand intercepts is the mean estimate (slightly) more than one ASE away from the true value. This is never the case for the household size or price coefficients. The mean estimate is more than one ASE from the true value for 8 out of 20 of the Cholesky parameters, but differs by (slightly) more than 2 ASE in only 2 cases. By this

[21] Note that, when dealing with MNP models with fairly large J, it is always advisable to estimate the Cholesky elements rather than elements of Σ. If one iterates on elements of Σ itself, it is quite common for them to move outside the range of feasible values for a positive definite covariance matrix during the course of the iterations. Iteration on the Cholesky elements guarantees that this cannot happen.

Table 4.1
Simulation estimation for MNP model[a]

θ	DGP	$\bar{\bar{\theta}}$	RMSE	$\overline{\text{ASE}}$	t-Bias	Bias/$\overline{\text{ASE}}$
β_{11}	−.307	−.319	0.110	0.091	−0.77	−0.09
β_{21}	−.961	−1.071	0.228	0.182	−3.41*	−0.60
β_{31}	0.163	−.034	0.289	0.183	−4.82*	−1.08*
β_{41}	−.946	−1.591	1.044	0.591	−4.37*	−1.09*
β_{51}	1.402	1.226	0.308	0.240	−4.04*	−0.73
β_{61}	0.954	0.888	0.118	0.097	−3.95*	−0.68
β_{12}	−.033	−.029	0.018	0.018	1.57	0.22
β_{22}	−.011	−.010	0.032	0.028	0.22	0.04
β_{32}	−.040	−.019	0.035	0.028	4.24*	0.75
β_{42}	−.035	0.000	0.068	0.047	3.64*	0.74
β_{52}	−.359	−.398	0.097	0.088	−2.84*	−0.44
β_{62}	−.171	−.171	0.022	0.025	0.00	0.00
γ	−1.981	−1.997	0.122	0.118	−0.92	−0.14
a_{21}	0.615	0.541	0.214	0.143	−2.45*	−0.52
a_{22}	1.019	1.032	0.125	0.115	0.74	0.11
a_{31}	0.410	0.457	0.195	0.147	1.70	0.32
a_{32}	0.443	0.339	0.256	0.163	−2.87*	−0.64
a_{33}	1.407	1.299	0.189	0.150	−4.04*	−0.72
a_{41}	0.322	0.530	0.453	0.244	3.25*	0.85
a_{42}	0.401	0.324	0.420	0.254	−1.30	−0.30
a_{43}	1.483	0.849	0.729	0.268	−6.15*	−2.37*
a_{44}	1.096	1.457	0.551	0.313	4.63*	1.15*
a_{51}	0.351	0.293	0.229	0.190	−1.79	−0.31
a_{52}	0.024	0.107	0.337	0.200	1.74	0.42
a_{53}	0.497	0.238	0.406	0.202	−4.51*	−1.28*
a_{54}	0.238	0.082	0.286	0.187	−3.86*	−0.83
a_{55}	0.905	1.004	0.284	0.249	2.46*	0.40
a_{61}	0.506	0.510	0.166	0.113	0.17	0.04
a_{62}	0.641	0.368	0.363	0.137	−5.32*	−1.99*
a_{63}	0.955	0.726	0.282	0.130	−5.74*	−1.76*
a_{64}	0.361	0.197	0.235	0.126	−4.93*	−1.30*
a_{65}	−.100	0.053	0.194	0.121	5.58*	1.26*
a_{66}	0.845	1.062	0.247	0.101	6.21*	2.15*

[a] Abbreviations: $\bar{\bar{\theta}}$, the mean parameter estimate across the 50 replications; RMSE, the root mean square error of the estimates about the true value; $\overline{\text{ASE}}$, the mean of the asymptotic standard errors across the 50 replications; t-Bias, the t-statistic for the bias, defined as $(\bar{\bar{\theta}} - \text{DGP})/(\text{RMSE}/\sqrt{50})$.

standard, one would have to conclude that SML based on GHK is delivering reasonably accurate estimates in a very difficult example[22].

Do the biases we observe in Table 4.1 stem primarily from the fact that simulation error enters the likelihood nonlinearly? To address this question we turn to the method of simulated moments (MSM) estimator developed in McFadden (1989). The solution of the simulated moment conditions:

$$\sum_{i=1}^{N}\sum_{j=1}^{J} W_{ij}\left[\chi\left[d_j=j\right]-\hat{p}\left(j|Z_i,X_i,\hat{\theta}\right)\right]=0, \tag{4.4}$$

where W_{ij} is a set of instruments and $\hat{p}$ is any unbiased simulator (satisfying certain regularity properties), gives an estimator for θ that is root N consistent for *fixed* simulation size M. The reason is that the simulation errors enter the moment conditions linearly, and hence tend to cancel across agents as $N \rightarrow \infty$ (provided that the simulation errors are asymptotically uncorrelated with the instruments). Geweke, Keane and Runkle (1994) also applied MSM to the same 50 artificial data sets considered in Table 4.1. The estimated biases were very similar for $\hat{\theta}_{\text{MSM}}$. In fact, the t tests indicated significant bias for 21 out of the 22 parameters for which $\hat{\theta}_{\text{SML}}$ had significant bias. Based on this, we conclude that the main source of bias in the SML estimates is small sample bias, rather than simulation induced bias. This statement may seem surprising given that $N = 5000$, but in fact such a sample size is not very large for a discrete choice model with 7 alternatives [Geweke et al. (1994) report that for artificial data sets with $N = 1000$ the various algorithms would often fail to converge due to numerical problems][23].

Based on these results, as well as results from a number of other numerical experiments in Geweke et al. (1994) and in other papers cited earlier, we conclude that in many contexts the finite sample bias in SML estimates is no more severe than for MSM estimates – provided that choice probabilities are simulated using the GHK method with a sufficiently large simulation size (apparently in the range of $M = 20$ to 50 in most contexts).

The original hope with MSM was that it would provide an inexpensive way to obtain MNP estimates with good statistical properties, because its simulation error cancellation property would allow one to rely on quite crude and inexpensive probability simulators. The Monte-Carlo literature has not born this out, because the finite sample performance of MSM appears to deteriorate quite substantially when

[22] Also encouraging is that there is fairly good agreement between empirical RMSE and the ASE's, at least for the β and γ parameters. There is some tendency for SML to understate standard errors, and this is much more severe for the Cholesky elements.

[23] Geweke et al. (1994) also applied a Gibbs sampling-data augmentation algorithm to the same data. If we treat the posterior means as Bayes estimates of the model parameters, estimated bias is in the same direction as for SML for 25 out of 33 parameters. And if the empirical standard deviation of the Bayes estimates is used to form t-statistics for the bias, it is significant in 11 out of 33 cases.

imprecise simulators are used. For instance, Geweke et al. (1994) found that MSM based on kernel smoothed (KS) frequency simulators [discussed in McFadden (1989)] produced estimates with much higher RMSEs than either SML–GHK or MSM–GHK. This was true even when the number of draws used to form the KS simulators was increased sufficiently to equate computation time with GHK, and regardless of which of several values of the smoothing parameter was employed. This led them to conclude "... the choice between estimation methods (i.e., MSM versus SML) is of secondary importance relative to the choice of probability simulator ..."

A related point concerns the fact that the optimal weights for MSM are of the form:

$$W_{ij}^{0} = \left[\partial p(j|Z_i, X_i, \theta)/\partial\theta\right]/p(j|Z_i, X_i, \theta). \tag{4.5}$$

The desirable consistency property of MSM obtains so long as one uses weights $\{W_{ij}\}$ that are asymptotically correlated with the $\left\{W_{ij}^{0}\right\}$ and uncorrelated with the simulated moments. But numerical experiments in Hajivassiliou (1991) indicate that MSM performs quite poorly unless reasonably accurate approximations to the optimal weights are employed. This implies that one needs to simulate the probabilities in Equation (4.5) to a reasonably high degree of accuracy.

There is one context where MSM has been shown to be clearly superior to SML, and that is in panel data models with serially correlated errors. The MSM estimator as originally formulated in McFadden (1989) is not practical in the panel data case, because each possible choice sequence must be treated as a separate alternative, meaning that J in Equation (4.4) grows exponentially with T. Keane (1990, 1993, 1994) suggested the alternative of factoring the sequence probabilities into transition probabilities to obtain the simulated moment conditions

$$\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{j=1}^{J} W_{itj}\left[\chi\left[d_{it}=j\right]-\hat{p}\left(j|D_{i,t-1}, Z_i, X_i, \hat{\theta}\right)\right]=0,$$

where J is the number of choice options in each period, $D_{i,t-1}$ is the history of choices $\{d_{i1}, \ldots, d_{i,t-1}\}$, and X_i and Z_i are the histories of the X_{it} and Z_{itj} variables. The optimal weights are the obvious generalization of Equation (4.5). Keane further proposed that the transition probabilities be simulated using ratios of GHK simulators:

$$\hat{p}\left(j|D_{i,t-1}, Z_{it}, X_{it}, \hat{\theta}\right) = \frac{\hat{p}_{\text{GHK}}\left(j, D_{i,t-1}|Z_{it}, X_{it}, \hat{\theta}\right)}{\hat{p}_{\text{GHK}}\left(D_{i,t-1}|Z_{it}, X_{it}, \hat{\theta}\right)}. \tag{4.6}$$

The resulting estimator is not consistent in N for fixed simulation size M because the denominator in Equation (4.6) is simulated. Hence simulation error does not enter the objective function linearly, and one must have $M/N^{1/2} \to \infty$ as $N \to \infty$ for this estimator to be $\sqrt{N}$ consistent and asymptotically normal (with a properly centered

Table 4.2
MSM and SML estimation for the multiperiod probit model[a]

θ	DGP	MSM			SML		
		$\bar{\bar{\theta}}$	RMSE	$\overline{\text{ASE}}$	$\bar{\bar{\theta}}$	RMSE	$\overline{\text{ASE}}$
ρ_1	0.800	0.808	0.022	0.018	0.755	0.048	0.014
ρ_2	0.800	0.790	0.042	0.037	0.680	0.124	0.027
a_{12}	0.500	0.567	0.121	0.107	0.620	0.138	0.067
a_{22}	0.866	0.892	0.109	0.097	1.012	0.161	0.064
β_{11}	0.500	0.500	0.033	0.042	0.499	0.030	0.033
β_{21}	−1.200	−1.176	0.090	0.084	−1.218	0.062	0.070
β_{12}	1.000	0.985	0.028	0.037	0.990	0.023	0.030
β_{22}	1.000	0.989	0.061	0.056	1.003	0.057	0.046
γ	1.000	0.982	0.038	0.036	0.991	0.025	0.030

[a] Abbreviations: $\bar{\bar{\theta}}$, the mean parameter estimate across the 20 replications; RMSE, the root mean square error of the estimates about the true value; $\overline{\text{ASE}}$, the mean of the asymptotic standard errors across the 20 replications.

limiting distribution). Nevertheless, this estimator has been found in a number of experiments to have finite sample properties that are superior to SML. For instance, Geweke, Keane and Runkle (1997) consider a three alternative model of the form:

$$y_{itj} = \beta_{j1} + \beta_{j2} X_{it} + \gamma Z_{itj} + \varepsilon_{itj} \qquad j = 1, 2,$$

and the normalization $y_{it3} = 0$, and with the error structure:

$$\varepsilon_{itj} = \rho_j \varepsilon_{i,t-1,j} + v_{itj} \qquad j = 1, 2,$$

where $(v_{it1} v_{it2})' \sim N(0, \Sigma)$. The element a_{11} of the Cholesky decomposition of Σ is fixed at 1, while a_{12} and a_{22} are estimated. Twenty artificial data sets of size $N = 500$ and $T = 10$ were constructed, for each of twelve alternative configurations of the true parameter vectors and the serial correlation structure of the covariates [we refer the reader to Geweke et al. (1997) for more details]. In Table 4.2 we present a representative set of their results (extracted from their Tables 16 and 17). These results are for MSM and SML each based on GHK with $M = 20$.

The severe downward bias of the SML estimates of the AR(1) parameters is apparent. The SML estimates of the cross-correlation parameters a_{12} and a_{22} are also severely biased. Simulation size had to be increased to about $M = 80$ or 160 before the biases became negligible. In contrast, the performance of MSM in this example is quite impressive, and it continued to produce good results even with $M = 10$. A similar pattern holds across all the experiments in Geweke et al. (1997). The pattern of SML

Table 4.3
SML estimates of Markov model[a]

T	M	ρ		$\lambda = 0.2$		$\beta = 1.0$	
		Mean	RMSE	Mean	RMSE	Mean	RMSE
$\rho = .40$							
15	50	.383	.042	.214	.055	.994	.044
30	50	.380	.034	.213	.037	.992	.030
50	50	.371	.035	.216	.032	.988	.027
$\rho = .85$							
15	50	.837	.024	.214	.070	.979	.055
30	50	.820	.034	.232	.057	.951	.062
50	50	.798	.054	.252	.064	.912	.093
15	15	.812	.045	.236	.078	.948	.072
30	15	.789	.063	.263	.081	.908	.099
50	15	.762	.089	.290	.099	.862	.134

[a] Abbreviations: mean, the mean estimate across the 300 replications; RMSE, the root mean square error of the estimates around the true values.

(but not MSM) severely underestimating serial correlation parameters was also found in Keane (1994).

Lee (1997) presents an extensive series of Monte-Carlo experiments in which SML based on GHK is applied to all the various dynamic panel data models presented in Heckman (1981). He finds that SML produces severely biased estimates of the degrees of serial correlation, heterogeneity and state dependence in many of these models. For example, Lee considers the Markov model with $J = 2$ and AR(1) errors:

$$y_{it1} = \beta X_{it} + \lambda\chi\left[d_{i,t-1} = 1\right] + \varepsilon_{it} \qquad t = 1, \ldots, T,$$

with $y_{it2} = 0$, $\varepsilon_{it} = \rho\varepsilon_{i,t-1} + \eta_{it}$, $\eta_{it} \sim N(0,1)$, $\varepsilon_{i0} = 0$ and $\chi\left[d_{i0} = 1\right] = 0$. He generates 300 artificial data sets with $N = 200$, and experiments with T, M, and the size of ρ. Table 4.3 contains a representative set of his results (extracted from his Table 2). Notice there is a consistent pattern of downward bias in ρ and upward bias in λ. These biases are more severe when ρ is high (.85) than when it is low (.40). The bias also becomes more severe as T increases. Also notice, however, that the bias is greatly reduced when M is increased from 15 to 50. But bias with $M = 50$ remains severe in the high serial correlation case, consistent with the Geweke et al. (1997) results.

Lee (1995) also proposed a bias reduction scheme based on approximating the leading bias term in the asymptotic expansion of the SML estimator that arises due to simulation. He implements this in Lee (1997) and concludes that “the bias-correction procedure reduces bias and RMSE, but the improvements ... are generally small.” Further, it appears that the bias correction works best when the initial bias in SML is not too great. This makes sense, since the accuracy of the asymptotic expansion will tend to be poor if one is far from the true θ. It appears that in cases of strong serial correlation it is more promising to simply increase simulation size [i.e., to say $M = 80$ to 160 as in Geweke et al. (1997)] than to attempt bias correction.

In summary, we conclude that simulation based estimation of MNP models, as well as other discrete choice models of similar complexity, is quite feasible using currently available methods. Gauss code to implement the both SML and MSM estimation of cross sectional and panel MNP models is available at the web site http://research.mpls.frb.fed.us/˜drunkle/software/GKR/mmp.html. But before attempting to use simulation based methods to estimate MNP models, it is important that several cautions and caveats be born in mind:

(1) In most contexts the choice between SML or MSM as the estimation method is not important (the one known exception being the case of panel data models with serially correlated errors, where the performance of MSM appears to be superior);

(2) but it is essential to use a highly accurate smooth probability simulator to implement either SML or MSM. The GHK method appears to be the most accurate general purpose simulator among those currently available.

(3) Care must be taken to use sufficiently large simulation sizes. For small sized problems (e.g., pure cross section problems with three or four alternatives) experience suggests that GHK based on only 10 to 30 draws will work well. But much larger simulation sizes are often necessary in more complex models. Geweke, Keane and Runkle (1994, 1997) and Lee (1995, 1997) give some guidance along these lines. In all cases it is advisable to check sensitivity of results to moderate increases in the number of draws.

(4) In specification of MNP models it is essential to bear in mind that alternative specific covariates (i.e., exclusion restrictions) are critical if covariance parameters are to be identified [see Keane (1992)]. Even so, covariance matrix parameters will often be poorly identified in models that have large choice sets. Furthermore, the number of covariance parameters grows rapidly as choice set size increases, and nonlinear search algorithms typically have difficulty with high dimensional parameter vectors. Hence, estimation of models with large choice sets will only be feasible if constraints are placed on the covariance matrix to reduce the number of free parameters. It is important to recognize that the difficulties described here are inherent to the MNP model itself, and are not a consequence of use of simulation methods per se[24].

[24] This point may seem obvious, but we have received many inquiries from investigators who concluded that simulation methods “don’t work” because they could not succeed in estimating a MNP model that lacked alternative specific covariates. Or because they found it impossible to estimate very high dimensional MNP models with hundreds of free covariance matrix elements.

5. Univariate latent linear models

Economic models are often used to study a single decision or outcome. The outcome variable may be fully observed, continuous, and unrestricted (for example, log consumption); fully observed and continuous but restricted to an interval (fraction of expenditure devoted to a certain category of goods); continuous but censored (earnings subject to known withholding limits for social insurance); a mixture of discrete and continuous outcomes (earnings of full-time high school students); categorical (income from survey data known only to be in a designated interval); or discrete (dichotomous choice, such as labor force participation). Depending on the model and data, other kinds of outcomes may be observed as well.

In all of these models, it is useful to conceive first of a latent outcome (denoted $\tilde{y}_t$, for observation t), and then a corresponding set-valued observed outcome, denoted y_t. For example, in the case of a continuous outcome censored from above at c, $y_t = \tilde{y}_t$ if $\tilde{y}_t \leqslant c$, and $y_t = (c, \infty)$ if $\tilde{y}_t > c$. In the case of a dichotomous outcome, one observes $y_t = (-\infty, 0]$ if $\tilde{y}_t \leqslant 0$ and $y_t = (0, \infty)$ if $\tilde{y}_t > 0$. This construction is sometimes used explicitly in introducing the tobit model [Amemiya (1985, Section 10.2) or Greene (1997, Section 20.3.2)] and probit model [Goldberger (1991, Section 29.1) or Maddala (1992, Section 8.9)], respectively.

This section treats the linear model $\tilde{y}_t = \beta' \boldsymbol{x}_t + u_t$, with observed outcomes of the form $y_t = [c_t, d_t], y_t = [c_t, d_t)$, or $y_t = (c_t, d_t]$, it being understood that $c_t \leqslant d_t$ and that c_t and d_t are extended real numbers. The disturbances u_t $(t = 1, \ldots, T)$ are independent and identically distributed conditional on $\boldsymbol{x}_t$ $(t = 1, \ldots, T)$. The disturbance u_t has a normal mixture distribution. We make this assumption because the normal mixture density can approximate any density arbitrarily well [Ferguson (1983)], and because it leads to practical methods for inference. It avoids the well-known problems that arise if the distribution of u_t is assumed to be Gaussian when in fact this assumption is poor. In the specific case of dichotomous choice models the strategy here has objectives similar to those of nonparametric single-index models[25]. The treatment here differs in that it covers a much wider class of latent variable models, is fully Bayesian, and is computationally less demanding than methods for single-index models.

Section 5.1 presents an overview of the univariate latent linear model (ULLM), leaving technical detail to Appendix A. Section 5.2 provides some results with artificial data, to establish the practicality of the methods. The ULLM is incorporated in the Bayesian Analysis, Computation and Communication (BACC) software system. This system provides extensions to Gauss, Matlab, and S-plus by means of dynamically linked libraries, making it easy for one familiar with one of these

[25] See for example Cosslett (1983), Manski (1985), Gallant and Nychka (1987), Powell, Stock and Stoker (1989), Horowitz (1992), Ichimura (1993), Klein and Spady (1993) and Lewbel (1997). For a detailed discussion of the use of mixture of normal models as an alternative to the probit model, see Geweke and Keane (1999).

commercial software packages to apply the model. Detailed information is available at http://www.econ.umn.edu/~bacc.

5.1. An overview of the univariate latent linear model

5.1.1. Distribution of disturbances

In the univariate latent linear model

$$\tilde{y}_t = \beta' \boldsymbol{x}_t + u_t, \tag{5.1}$$

the disturbances u_t ($t = 1, \ldots, T$) are i.i.d. conditional on $\boldsymbol{x}_t$ ($t = 1, \ldots, T$). Several alternative assumptions about the distribution of u_t can be made, and here we shall take up three in detail. The first is the conventional specification $u_t \sim N(0, \sigma^2)$, in which σ^2 may be a free parameter (for example, in the censored linear model) or fixed as a condition of identification (for example, $\sigma^2 = 1$ in the probit model).

The second alternative assumption about the distribution is $u_t \sim t(0, \sigma^2; \lambda)$, a Student-$t$ distribution with location parameter 0, scale parameter σ, and degrees-of-freedom parameter λ. The scale parameter may be fixed as a condition of identification. The disturbances may be represented $u_t = \sigma_{(t)} \eta_t$, with $(\sigma_{(1)}, \ldots, \sigma_{(T)})$ and $(\eta_1, \ldots, \eta_T)$ i.i.d. and mutually independent conditional on $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T)$. The latent variables $\sigma^2_{(t)}$ have independent inverted gamma distributions, $\lambda / \sigma^2_{(t)} \sim \chi^2(\lambda)$, and $\eta_t \sim N(0, \sigma^2)$[26]. They subsequently play an important part for inference in this model.

The third alternative assumption about the distribution is $u_t \sim N(\alpha_j, \sigma^2 \sigma_j^2)$ with probability p_j ($j = 1, \ldots, m$); $\sum_{j=1}^m p_j = 1$. This is a normal mixture model, with u_t drawn at random from one of m "urns", each urn containing a collection of u_t with a different normal distribution. By increasing the value of m and choosing the $N(\alpha_j, \sigma^2 \sigma_j^2)$ distributions appropriately, any univariate p.d.f. can be approximated arbitrarily well in the L_1 topology [Ferguson (1983)]. In this case, the disturbance may be represented $u_t = \alpha' \tilde{\boldsymbol{z}}_t + \sigma_{(t)} \eta_t$. In this representation $\alpha' = (\alpha_1, \ldots, \alpha_m)$. The random variables $(\eta_1, \ldots, \eta_T)$ are i.i.d. conditional on $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T)$: $\eta_t \sim N(0, \sigma^2)$. The latent random vectors $(\tilde{\boldsymbol{z}}_t', \sigma_{(t)})$ are i.i.d. conditional on $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T)$ and $(\eta_1, \ldots, \eta_T)$. Their values are governed by a latent state variable $s(t)$ taking on the alternative values $s(t) = j$ ($j = 1, \ldots, m$). The $s(t)$ are i.i.d. conditional on $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T)$ and $(\eta_1, \ldots, \eta_T)$, with $P[s(t) = j] = p_j$. Conditional on $s(t) = j$, we have $\sigma_{(t)} = \sigma_j$, $\tilde{z}_{tj} = 1$, and $\tilde{z}_{ti} = 0 \, \forall \, i \neq j$. To identify the model with respect to permutation of the state index, it is assumed that $\sigma_1 > \cdots > \sigma_m$. Identification of σ separately from σ_j ($j = 1, \ldots, m$) is taken up subsequently as part of the prior distribution.

[26] For a derivation of this construction see Johnson, Kotz and Balakrishnan (1995, Section 28.1) or Geweke (1993).

All three specifications of the distribution of u_t in Equation (5.1) are embedded in

$$\tilde{y}_t = \alpha' \underset{m\times 1}{\tilde{\boldsymbol{z}}_t} + \beta' \underset{k\times 1}{\boldsymbol{x}_t} + \varepsilon_t, \qquad \varepsilon_t = \sigma_{(t)}\eta_t,$$

in which, conditional on $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T)$, $\eta_t \sim N(0, \sigma^2)$ is i.i.d. and independent of $(\tilde{\boldsymbol{z}}_t', \sigma_{(t)})$. The three specifications of the distribution of u_t in Equation (5.1) are distinguished by the distribution of $\sigma_{(t)}$. Only when u_t has a normal mixture distribution is $m > 0$.

5.1.2. *Observable outcomes*

Models are further distinguished by the observable outcome y_t, which in general is a set-valued function of the latent outcome $\tilde{y}_t$. If $y_t = \tilde{y}_t$ $(t = 1, \ldots, T)$ the ULLM reverts to the linear model. For the dichotomous choice model $y_t = (-\infty, 0]$ if $\tilde{y}_t \leqslant 0$ and $y_t = (0, \infty)$ if $\tilde{y}_t > 0$. For an outcome censored from above at c, $y_t = \tilde{y}_t$ if $\tilde{y}_t \leqslant c$, and $y_t = (c, \infty)$ if $\tilde{y}_t > c$. In all cases,

$$p(y_t, \tilde{y}_t | \boldsymbol{x}_t) = p(\tilde{y}_t | \boldsymbol{x}_t)\, p(y_t | \tilde{y}_t) = p(\tilde{y}_t | \boldsymbol{x}_t)\, \chi_{y_t}(\tilde{y}_t),$$

in which $\chi_S(z)$ is the set indicator function: $\chi_S(z) = 1$ if $z \in S$ and $\chi_S(z) = 0$ if $z \notin S$.

5.1.3. *Prior distributions*

Every model is endowed with a proper prior distribution. This makes it possible to compare different models for the same data using Bayes factors, as discussed below. For each model, we specify a benchmark prior distribution with hyperparameters that can be adjusted to reflect beliefs. These prior distributions are chosen for their combination of flexibility and analytical simplicity. Beyond the choice of hyperparameters, these prior distributions may be adjusted further to include prior distributions not in the benchmark families, by reweighting the output of the posterior simulator constructed subsequently[27].

The benchmark prior distribution for β is Gaussian, $\beta \sim N(\underline{\beta}, \underline{\boldsymbol{H}}_\beta^{-1})$, and for σ^2 it is inverted gamma, $\underline{s}^2/\sigma^2 \sim \chi^2(\underline{\nu})$. If $\underline{s}^2/\underline{\nu} = \sigma^{*2}$ and $\underline{s}^2 \to \infty$, then σ^2 is degenerate at σ^{*2}. Thus $\sigma^2 = 1$ can be enforced by a very large value of $\underline{s}^2 = \underline{\nu}$.

For the Student-t model the benchmark prior for the degrees of freedom parameter is exponential with mean $\underline{\lambda}$, $\lambda \sim \exp(\underline{\lambda})$. Smaller values of $\underline{\lambda}$ reflect beliefs that the distribution is more leptokurtic.

The normal mixture model for the disturbances has three components: $\boldsymbol{p}' = (p_1, \ldots, p_m)$, $(\sigma_1^2, \ldots, \sigma_m^2)$, and $\alpha' = (\alpha_1, \ldots, \alpha_m)$. The multinomial distribution of the

[27] Such reweighting is discussed in Geweke (1999, Section 6) and is easy to carry out in the BACC software system.

state index $s(t) = j(j = 1, \ldots, m)$ involves the probabilities $p_1, \ldots, p_m$, $\sum_{j=1}^{m} p_j = 1$. The benchmark prior distribution is Dirichlet (multivariate beta) with hyperparameters $r_1, \ldots, r_m$: $p(\boldsymbol{p}) \propto \prod_{j=1}^{m} p_j^{r_j - 1}$. In this distribution p_j has mean r_j / R and standard deviation $[r_j(R - r_j)]^{1/2} / R(R+1)^{1/2}$, where $R = \sum_{i=1}^{m} r_i$.

The benchmark prior distribution for the second component of the normal mixture model, the variance scaling parameters σ_j^2, consists of the m inverted gamma components $\underline{s}_j^2 / \sigma_j^2 \sim \chi^2(\underline{\nu}_j)$. These are subject to the restrictions $\sigma_1^2 > \cdots > \sigma_m^2$ but otherwise independent. The ordering of the σ_j^2 removes the possibility of permuting the states, but imposes the restriction $\sigma_i^2 \neq \sigma_j^2 \ \forall \ (i,j)$. We choose to identify states by variance, because in our applications this is more convenient than identifying states by orderings of state probabilities, p_j, or state means, α_j. The lack of identification in the likelihood imposed by the fact that the variance in component j is $\sigma^2 \sigma_j^2$ is resolved by the proper prior distributions for σ^2 and $(\sigma_1^2, \ldots, \sigma_m^2)$. Identification can also be achieved in the traditional way by taking $\underline{s}_i^2 \to \infty$ and $\underline{\nu}_i \to \infty$ while $\underline{s}_i^2 / \underline{\nu}_i = 1$, for a selected state i, thereby making σ^2 the variance in that state. In either event, in the prior distribution of variance ratios across states,

$$\frac{\sigma^2 \sigma_k^2}{\sigma^2 \sigma_j^2} \sim \frac{\underline{s}_k^2 / \underline{\nu}_k}{\underline{s}_j^2 / \underline{\nu}_j} \cdot F\left(\underline{\nu}_j, \underline{\nu}_k\right),$$

subject to $\sigma^2 \sigma_j^2 / \sigma^2 \sigma_k^2 < 1$ if $j > k$. Thus the prior distribution for the σ_j^2 incorporates only beliefs about relative variances. This is convenient in thinking about shapes (as opposed to scales) of distributions – for example, outliers or other forms of leptokurtosis.

The third component of the normal mixture distribution, $\alpha' = (\alpha_1, \ldots, \alpha_m)$, is multivariate normal, $\alpha | \sigma \sim N(\mathbf{0}, \sigma^2 \underline{\boldsymbol{H}}_\alpha^{-1})$. This prior distribution is taken conditional on σ, and prior variance is proportional to σ^2, in order to represent beliefs about the shape of the disturbance p.d.f. independent of its scale. In all of the illustrations in the next subsection, $\underline{\boldsymbol{H}}_\alpha = \underline{h}_\alpha \boldsymbol{I}_m$. This restriction requires these prior beliefs about means to be interchangeable across the mixture components. For example, given the number of states, m, the greater the precision $\underline{h}_\alpha$, the more likely is the p.d.f. to be unimodal.

5.1.4. Existence of the posterior

For any particular ULLM the product of the relevant prior densities and data density is the kernel of the posterior distribution, so long as that product is finitely integrable over the space of all parameters and latent variables. If the data density is bounded above, then the integrability condition is met when the prior distribution is proper (as it is here). For all variants of the ULLM with normal and Student-t densities the data density is bounded, as it is for all variants in which all outcomes are discrete (i.e., $c_t < d_t \ \forall \ t$). If the disturbances are mixed normal and at least some of the $\tilde{y}_t$ are not latent (i.e., $c_t = d_t$ for at least some t) then the data density is unbounded. In this case the integrability of the posterior kernel can be demonstrated, but the argument is more technical and is relegated to Appendix A.

5.1.5. MCMC algorithm for inference

The explicit development of the data density and prior density, whose product is the posterior density kernel, is given in Appendix A. There is no corresponding closed form for the posterior distribution of all of the parameters jointly. However, the Gibbs sampling algorithm described in Section 2.2 can be applied to eight groups of parameters or latent variables that appear in the posterior kernel: (α, β); σ^2; $\sigma^2_{(t)}$ $(t = 1, \ldots, T)$; λ; $s(t)$ $(t = 1, \ldots, T)$ and $\tilde{\boldsymbol{Z}}$; $\boldsymbol{p}$; σ^2_j $(j = 1, \ldots, m)$; and $\tilde{y}_t$ $(t = 1, \ldots, T)$. (Not all parameters appear under each assumption about the distribution of u_t.)

The Gibbs sampling algorithm is practical because the distribution of each parameter group, conditional on all the others, is simple enough that draws from the conditional distribution can be made. In particular, the conditional distribution of (α, β) is multivariate normal; those of σ^2 and $\sigma^2_{(t)}$ $(t = 1, \ldots, T)$ are inverted gamma; and each $\tilde{y}_t$ is truncated normal. In the Student-t model, the conditional distribution of λ is not standard, but a Metropolis within Gibbs step (Section 2.5) employing a Gaussian approximation to the conditional distribution as the proposal distribution works well. (Details are presented in Appendix A.) In the normal mixture model, the conditional distribution of$\boldsymbol{p}$ is Dirichlet and the state assignments are multinomial. The conditional distribution of σ^2_j is inverted gamma, subject to truncation restrictions imposed by the ordering $\sigma_1 > \cdots > \sigma_m$.

For seven of the eight groups of parameters in this algorithm, the support of the conditional posterior distribution is the same as the support of the marginal posterior distribution. The exception is the group of variance parameters σ^2_j $(j = 1, \ldots, m)$ in the normal mixture model. Thus for the normal and Student-t variants of the ULLM, Corollary 2.4.2 assures convergence of the Gibbs Markov chain to the posterior distribution.

For the normal mixture model, consider any point in the support of the posterior density, and any subset of the posterior density support with positive posterior probability. There exists a finite number of iterations of the algorithm such that the transition probability from the point to the subset is positive. (The minimum number of steps will depend on the values of the σ^2_j $(j = 1, \ldots, m)$ at the point in question, and their values in the subset. In the normal and Student-t models, this minimum number of steps is one for any point and any subset combination.) This condition assures that the transition density of the chain is aperiodic and absolutely continuous with respect to the posterior density [Tierney (1994)]. From Corollary 1 of Tierney (1994), the sequence of parameters and latent variables $\{\theta^{(m)}\}$ produced by the Markov chain is ergodic: that is, if a posterior moment $\bar{g} = E[g(\theta)|\boldsymbol{X}, \boldsymbol{y}]$ exists, then $\bar{g}_M = M^{-1} \sum_{m=1}^{M} g(\theta^{(m)}) \xrightarrow{a.s.} \bar{g}$.

5.1.6. Marginal likelihoods

It is useful to be able to compare two alternative specifications of the ULLM – for example, models with different specifications of the disturbance distribution, with different covariates $\boldsymbol{X}$, or with different prior distributions. A formal comparison can be made by means of a posterior odds ratio. Let A_1 and A_2 denote the alternative specifications of the ULLM and $p(A_1)$ and $p(A_2)$ the prior probabilities of the alternative model specifications themselves. Then the posterior odds ratio in favor of the specification A_1 is

$$\frac{p(A_1|\boldsymbol{X},\boldsymbol{y})}{p(A_2|\boldsymbol{X},\boldsymbol{y})}=\frac{p(A_1)\,p(\boldsymbol{y}|A_1)}{p(A_2)\,p(\boldsymbol{y}|A_2)},$$

in which the marginal likelihoods $p(\boldsymbol{y}|A_j)$ are given by Equation (2.2). The key technical task is to evaluate the integrals in Equation (2.2).

For the ULLM, a convenient way to approximate the marginal likelihood is to use the modified harmonic mean method of Gelfand and Dey (1994), as further developed in Geweke (1999, Section 4.3). This method requires that the prior densities $p(\theta_j^{(m)}|A_j)$ and data densities $p(\boldsymbol{y}|\boldsymbol{X}, \theta_j^{(m)}, A_j)$ be evaluated for each iteration m of the MCMC algorithm. Once this is done, the Monte Carlo approximation of the marginal likelihood may be carried out using generic software described in Geweke (1999, Section 4.5)[28]. The evaluation of the data density $p(\boldsymbol{y}|\boldsymbol{X}, \theta_j^{(m)}, A_j)$ is relatively straightforward in the ULLM, but some care is required in the handling of the latent variables. Details are given in Appendix A.

5.2. Some evidence from artificial data

Before proceeding to apply the ULLM, a number of practical issues arise. For some variants of the ULLM, it is simply of interest to see how well the posterior distribution recovers the underlying population parameters. This is especially true of models with latent $\tilde{y}_t$ – for example, what can be learned about the degrees of freedom parameter λ in the Student-t dichotomous choice model, or the conditional means α_j in the censored linear model or dichotomous choice model with mixed normal disturbances?

In all cases, it is important to ascertain some information about computational efficiency. This is not simply a matter of the computation time required for each iteration of the MCMC algorithm. It is also driven by the degree of serial correlation of the parameters drawn from one iteration to the next. The variance of the numerical approximation of the posterior mean is computed using conventional time series

[28] Other methods for Monte Carlo approximation include Chib (1995) and importance sampling as discussed in Geweke (1996). Neither applies directly to a Gibbs sampling algorithm with Metropolis steps.

methods. The method is, essentially, to apply a set of linearly declining weights to the first 8% of the autocovariances of the sequence $\{g(\theta^{(m)})\}$, for a given function of interest $g(\cdot)$[29].

5.2.1. Parameter posterior moments

Tables 5.1–5.3 provide model specifications, prior distributions, and some posterior moments for instances of the univariate linear model, censored linear model, and dichotomous choice linear model, respectively. In each case sample size is $T = 2000$, $x_{t1} = 1.0$ is an intercept, and x_{2t} and x_{3t} are independent, i.i.d. standard normal variates. The nine data sets, including covariates, were drawn independently. The MCMC algorithm was executed for 12 000 iterations in each case, and the last $M = 10\,000$ iterations were used for the computations. Each table shows the parameter values used to generate the data.

Panel C of Table 5.1 provides posterior moments for the textbook normal linear model. There are no surprises: the posterior standard deviations of the β_j are all about $(2000)^{-1/2}$ and that of σ is about $(1000)^{-1/2}$, the values suggested by the design of the experiment. Given the approximate orthogonality of β and σ^2 in the posterior distribution the MCMC draws should be nearly i.i.d., and this is reflected in RNEs close to 1.0[30].

The Student-t linear model (Table 5.1, panel D) entails draws of $\sigma^2_{(t)}$ ($t = 1, \ldots, T$) each iteration. This accounts for the doubling of computation time per iteration, compared with the normal linear model. There is a modest increase in the posterior standard deviation of β and σ[31]. The posterior mean of the degrees of freedom parameter λ is close to the population value and well within one posterior standard deviation. Additional serial correlation (relative to the normal linear model) is introduced to the MCMC algorithm by the addition of $\sigma^2_{(t)}$ ($t = 1, \ldots, T$) and λ to the parameter list. This has at most a modest impact on the RNE of the approximation of $E(\beta_j|\boldsymbol{X},\boldsymbol{y})$. The main impact is on the numerical approximation error for σ. This arises because of the positive posterior correlation of σ and λ, and the fact that they are drawn in separate blocks of the MCMC algorithm. The numerical standard error of λ is reduced to 10% of its posterior standard deviation in about 3000 iterations, requiring about two minutes of computing time.

[29] For long simulations, care must be taken to achieve computational efficiency in the computation of autocovariances. Geweke (1999, Section 3.8) provides details.

[30] The same algorithm is applied to this simple normal linear model as is applied in the ULLM generally. Thus, moment matrices like $\boldsymbol{X}'\boldsymbol{X}$ are recomputed each iteration. Code designed specifically for the normal linear model, such as that in the BACC software (http://www.econ.umn.edu/~ bacc) is substantially more efficient.

[31] Note that the population variance of the disturbance in this model is $\sigma^2\lambda/(\lambda-2) = 5/3$. When the normal linear model is applied to this data set, the posterior standard deviation of the β_j is quite close to the value of $[(5/3)/2000]^{-1/2}$ that one would expect.

Table 5.1
Univariate linear model ($T = 2000$)

A. Model specification

$y_t = \beta_1 + \sum_{j=2}^{3} x_{jt} + u_t$

Population for all variants: $x_{2t} \overset{\text{i.i.d.}}{\sim} N(0,1)$, $x_{3t} \overset{\text{i.i.d.}}{\sim} N(0,1)$, $\beta_1 = 0$, $\beta_2 = 1$, $\beta_3 = -1$

Normal disturbances: $u_t \overset{\text{i.i.d.}}{\sim} N(0,1)$

Student-t disturbances: $u_t \overset{\text{i.i.d.}}{\sim} t(0,1;5)$

Mixed normal disturbances: $u_t \overset{\text{i.i.d.}}{\sim} N(-.3,1)$, $p_1 = .5$; $u_t \overset{\text{i.i.d.}}{\sim} N(.3,.2^2)$, $p_1 = .5$

B. Prior distributions and moments

Parameters	Prior distribution	Prior mean	Prior s.d.
$\beta_j (j = 1,2,3)$	$\beta_j \sim N(0,1)$	0.0	1.0
σ	$4/\sigma^2 \sim \chi^2(4)$	1.253	0.655
λ	$\lambda \sim \exp(5)$	5.0	3.162
$\alpha_j (j = 1,2)$	$\alpha_j \sim N(0,5\sigma^2)$	0.0	2.236σ
σ_j	$4/\sigma_1^2 \sim \chi^2(4)$	1.253	1.414
	$0.4/\sigma_2^2 \sim \chi^2(4)$	0.396	0.447
p_1	Beta(2,2)	0.500	0.224

Some posterior moments

Parameter	Mean	Stan. dev.	RNE	Parameter	Mean	Stan. dev.	RNE
C. Normal disturbances; .018 sec./iter.							
$\beta_1 = 0$	–.031	0.023	0.808	$\beta_3 = -1$	–.991	0.022	1.380
$\beta_2 = 1$	1.014	0.022	1.112	$\sigma = 1$	1.000	0.016	1.380
D. Student-t disturbances; .037 sec./iter.							
$\beta_1 = 0$	–.026	0.025	0.505	$\beta_3 = -1$	–1.016	0.024	1.052
$\beta_2 = 1$	0.968	0.025	0.538	$\sigma = 1$	0.971	0.029	0.063
				$\lambda = 5$	5.111	0.702	0.032
E. Normal mixture disturbances; .052 sec./iter.							
$\beta_2 = 1$	1.004	0.007	1.054	$\sigma \cdot \sigma_1 = 1$	0.965	0.022	0.745
$\beta_3 = -1$	–1.019	0.007	0.412	$\sigma \cdot \sigma_2 = .2$	0.193	0.008	0.007
$\beta_1 + \alpha_1 = -.3$	–.303	0.008	0.520	$p_1 = .5$	0.511	0.018	0.107
$\beta_1 + \alpha_2 = .3$	0.260	0.033	0.477				

The normal mixture model (Table 5.1, panel E) contains two normal components of the disturbance, each with the same probability but different means and variances. (The same mixture is used in the mixed normal censored linear and dichotomous choice models.) The variance of the disturbance is .52, smaller than in the normal models. If the normal linear model is applied to this data set, posterior standard deviations of β_2 and β_3 are about $.016 = (.52/2000)^{1/2}$ as one would expect. If the states were known, the posterior standard deviations of β_2 and β_3 would be about $.0062 = (1000/.2^2 + 1000/1^2)^{-1/2}$. The actual posterior standard deviations are much closer to the latter value than the former. That they exceed .0062 can be attributed to the imperfect sorting of observations by state. The posterior means of the intercept values are well within two posterior standard deviations of population values. Since $\sigma_1 = 5\sigma_2$, $\text{var}(\beta_1 + \alpha_1|\boldsymbol{X},\boldsymbol{y}) > \text{var}(\beta_1 + \alpha_2|\boldsymbol{X},\boldsymbol{y})$, but $\text{var}(\beta_1 + \alpha_1|\boldsymbol{X},\boldsymbol{y})/\,\text{var}(\beta_1 + \alpha_1|\boldsymbol{X},\boldsymbol{y}) < 5$ again due to imperfect sorting by states.

The low value of relative numerical efficiency indicates strong serial correlation in σ_2 in the MCMC algorithm. This arises because of high correlation between σ_2 and the state classifications $s(t)$. Because of the contrast in standard deviations ($\sigma_1 = 5\sigma_2$), there are only a few observations for which $p[\varepsilon_t|s(t) = 1] \approx p[\varepsilon_t|s(t) = 2]$, and therefore there is substantial persistence in state classification. Simple arithmetic shows that the reclassification of an observation has a much larger effect on the smaller variance. Since the effects of changes in other parameters on σ_1 and σ_2 is about the same, σ_2 is more strongly driven by the slowing moving state assignments.

Table 5.2 presents similar information for the censored linear model. In this model $\tilde{y}_t$ is observed if and only if $\tilde{y}_t > 0$, so about half of the $T = 2000$ observations are censored. Compared with the linear model, posterior standard deviations are in every case higher, reflecting the loss of information in censoring. Since about half the $\tilde{y}_t$ must be drawn each iteration, one would expect an increase in serial correlation. This is reflected in a reduction of RNE for all parameters except $\sigma \cdot \sigma_2$ in the normal mixture model. Comparisons among panels C, D and E in Table 5.2 are similar to the comparisons already discussed for their counterparts in Table 5.1. The increase in RNE for $\sigma \cdot \sigma_2$ is due to the fact that uncertainty about $\tilde{y}_t$ for those $\tilde{y}_t < 0$ now contributes in a major way to all parameters, and serial correlation in the draws of $\tilde{y}_t$ from one simulation to the next contributes to the serial correlation in all parameters in the MCMC algorithm. Thus, the contrast in the impact of state classification on σ_1 and σ_2 is less important, relative to all other factors contributing to serial correlation, than was the case in the linear model with mixed normal disturbances.

Table 5.3 presents the same information for the dichotomous choice linear model. Given the parameter values and the distribution of x_t, the probabilities of the choices are .5 for the normal and Student-t disturbances, and the probability of choice one ($\tilde{y}_t < 0$) is .47 for the mixed normal disturbances. With the obvious exception of the parameters σ and σ_1, which are normalized at 1.0, posterior standard deviations for all parameters are higher than was the case in the censored linear model. The largest increases are in the posterior standard deviations of the parameters of the disturbance distribution (other than σ and σ_1). The increase in the posterior standard deviation of the degrees of freedom parameter λ is especially large. Given the increased latency of $\tilde{y}_t$ in this model, these developments are unsurprising. There is evidence of increased serial correlation (relative to the censored linear model) in the MCMC algorithm, in

Table 5.2
Univariate censored linear model ($T = 2000$)

A. Model specification

$\tilde{y}_t = \beta_1 + \sum_{j=2}^{3} x_{jt} + u_t; y_t = \chi_{(0,\infty)}(\tilde{y}_t)\,\tilde{y}_t + \chi_{(-\infty,0)}(\tilde{y}_t)\,(-\infty, 0)$

Data generating process otherwise as in Table 5.1

B. Prior distributions and moments: as in Table 5.1

Some posterior moments

Parameter	Mean	Stan. dev.	RNE	Parameter	Mean	Stan. dev.	RNE
C. Normal disturbances; .044 sec./iter.							
$\beta_1 = 0$	–.048	0.034	0.101	$\beta_3 = -1$	–1.013	0.031	0.154
$\beta_2 = 1$	1.020	0.032	0.234	$\sigma = 1$	1.009	0.024	0.188
D. Student-t disturbances; .070 sec./iter.							
$\beta_1 = 0$	–.063	0.037	0.181	$\beta_3 = -1$	–1.036	0.034	0.167
$\beta_2 = 1$	1.030	0.036	0.262	$\sigma = 1$	0.943	0.040	0.017
				$\lambda = 5$	4.478	0.708	0.011
E. Normal mixture disturbances; .095 sec./iter.							
$\beta_2 = 1$	1.002	0.015	0.060	$\sigma \cdot \sigma_1 = 1$	0.978	0.031	0.672
$\beta_3 = -1$	–1.019	0.008	0.839	$\sigma \cdot \sigma_2 = .2$	0.197	0.021	0.107
$\beta_1 + \alpha_1 = -.3$	–.293	0.020	0.038	$p_1 = .5$	0.477	0.025	0.070
$\beta_1 + \alpha_2 = .3$	0.216	0.049	0.139				

the form of reduced RNEs, when panels C, D, and E in Table 5.3 are compared with their counterparts in Table 5.2. This decreased efficiency is most pronounced in the case of mixed normal disturbances. (Note that now the RNE of σ_2 is comparable with that of other parameters.)

Inference in the dichotomous choice linear model with mixed normal disturbances is reliable, in the sense that for all the parameters the posterior standard deviation is substantially less than the prior standard deviation, and all posterior means are within about one posterior standard deviation of the population values. There is very substantial serial persistence in the MCMC algorithm, but each iteration requires only about 0.1 seconds. Based on an RNE of .008, numerical standard errors are driven to one-fourth of posterior standard deviation after about 2000 iterations (about three minutes), to one-tenth after 12 500 iterations (about 20 minutes), and to 1% after 1.25×10^6 iterations (about 1.5 days). The practicality of the procedure thus depends on one's standards for accuracy. As we shall now see, a great deal can be learned with just a few thousand iterations.

Table 5.3
Dichotomous choice linear model ($T = 2000$)

A. Model specification

$\tilde{y}_t = \beta_1 + \sum_{j=2}^{3} x_{jt} + u_t; y_t = \chi_{(0,\infty)}\ (\tilde{y}_t)\ (0,\infty) + \chi_{(-\infty,0)}\ (\tilde{y}_t)\ (-\infty,0)$

Data generating process otherwise as in Table 5.1

B. Prior distributions and moments

Parameters[1]	Prior distribution	Prior mean	Prior s.d.
σ	$\sigma = 1$	1.0	0.0
σ_j	$\sigma_1 = 1$	1.0	0.0
	$0.4/\sigma_2^2 \sim \chi^2(4)$	0.396	0.447

Some posterior moments

Parameter	Mean	Stan. dev.	RNE	Parameter	Mean	Stan. dev.	RNE
C. Normal disturbances; .067 sec./iter.							
$\beta_1 = 0$	−.035	0.036	0.285	$\beta_3 = -1$	−1.086	0.050	0.096
$\beta_2 = 1$	1.040	0.049	0.555				
D. Student-t disturbances; .096 sec./iter.							
$\beta_1 = 0$	−.011	0.048	0.018	$\beta_3 = -1$	−1.283	0.147	0.004
$\beta_2 = 1$	1.265	0.141	0.004	$\lambda = 5$	3.639	1.808	0.003
E. Normal mixture disturbances; .139 sec./iter.							
$\beta_2 = 1$	1.032	0.080	0.003	$\sigma_2 = 0.2$	0.302	0.055	0.007
$\beta_3 = -1$	−1.051	0.080	0.004	$p_1 = .5$	0.462	0.080	0.004
$\beta_1 + \alpha_1 = -.3$	−.361	0.066	0.004				
$\beta_1 + \alpha_2 = .3$	0.395	0.129	0.011				

[1] $\beta_j (j = 1,2,3)$, λ, $\alpha_j (j = 1,2)$ and p_1 as in Table 5.1.

5.2.2. Marginal likelihood approximations

In the case of the ULLM, the additional computations required to produce the marginal likelihood are trivial. The likelihood function and prior distribution must be evaluated for those iterations that are used to approximate the marginal likelihood. Since these evaluations are not used subsequently in the MCMC algorithm they need not be made every iteration, but doing so in each iteration increases the computation time only by about 2%. Given the sequence of likelihood and prior evaluations from the MCMC algorithm, the log marginal likelihood is approximated using the variant of the Gelfand and Dey (1994) procedure described above. Computing time for this approximation, using the implementation detailed in Geweke (1999, Section 4.3), is

Table 5.4
Log marginal likelihoods in some univariate latent linear models with artificial data

Model	Data disturbances		
	Normal	Student-t	Mixed normal
A. Linear model			
Normal	–2851.4	–3268.2	–2297.7
Student-t	–2855.5	–3207.2	–2155.3
Mixed normal	–2857.8	–3213.5	–1900.4
B. Censored linear model			
Normal	–1821.5	–2037.4	–1575.1
Student-t	–1826.3	–2001.8	–1528.6
Mixed normal	–1826.5	–2003.3	–1403.3
C. Dichotomous choice model			
Normal	–802.1	–851.8	–677.4
Student-t	–804.5	–847.0	–673.7
Mixed normal	–807.8	–849.3	–662.4

essentially proportional to the number of iterations – about 2 seconds for 10 000 iterations.

Table 5.4 shows several patterns of results. First, for each data set (column headings) the model (row headings) that generates the observations receives the highest marginal likelihood, and therefore the highest posterior probability, of the three models compared. The lowest odds ratio in favor of any true model is that in favor of the Student-t over the mixed normal censored linear model, about 4.5:1. Second, the highest odds ratios occur when the competitor to a true model does not nest or approximate the true model – that is, Student-t versus normal when disturbances are Student-t, and normal mixture versus either normal or Student-t when the disturbances are normal mixture. Third, odds ratios are usually higher when the outcome ($\tilde{y}_t$) is fully observed than when it is not: they tend to be highest in the linear model, lowest in the dichotomous choice model. Fourth, odds ratios in favor of true models against nesting models (e.g., in favor of normal versus Student-t when disturbances are normal) or in favor of true models against approximating models (e.g., Student-t versus normal mixture when disturbances are Student-t) are lower, and the degree of latency has little effect on the magnitude of the odds ratio.

We conclude that in these examples, discrimination between the three disturbance distributions is effective. Moreover while the results shown in Table 5.3 are based on 10 000 MCMC iterations after discarding the first 2000, nearly identical results are obtained with 900 iterations after discarding the first 100. For these examples, an investigator could learn quickly which models account for most of the posterior probability and then concentrate computing resources on those models.

5.3. Some evidence from real data

Especially in large data sets, it is relatively easy to detect departures from normality and establish the form of the non-Gaussian distribution with a high degree of precision. Space constraints do not permit development of these applications in detail, so we confine this discussion to three examples in our recent work.

Geweke and Keane (2000) estimates a reduced form life cycle earnings model of the kind introduced by Lillard and Willis (1978) and used in a succession of studies since. A recurring puzzle in this literature has been the inability of these models to capture the transition of individual earnings in and out of the lowest quintile of the earnings distribution. Geweke and Keane (2000) adopts the standard model, but departs from it in two specific ways. First and most important, it specifies shocks to current earnings to be a mixture of three normals. Second, it sets up the regression of earnings on age and education as a high-order polynomial. (There are other elaborations as well, but these are the most important in the context of the ULLM.) Among the paper's many findings, three are important with respect to the ULLM. First, the evidence against normality is overwhelming: the distribution of the shock to current earnings is strongly skewed and leptokurtic. When the same model is fit using a normal distribution, the .40 quantile of the fitted normal corresponds to the .20 quantile of the mixture of three normal distributions. Second, the normal mixture model implies dynamics for the movement of individual earnings in and out of the lowest quintile that are quite similar to those in the data, and much closer than has been captured previously by reduced form life cycle earnings models in the literature. Third, maximum likelihood estimation in this model is not possible, because the likelihood function has a multitude of isolated singularities. [This point is discussed briefly in Appendix A of this chapter, and in greater detail in on-line Appendix F of Geweke and Keane (2000)].

Two simpler applications appear in Geweke, McCausland and Stevens (2000) and Geweke and Keane (1999). The former example is a simple hedonic regression model for residential real estate prices. The sample size is modest ($n = 546$) and the departure from normality is small (least squares residual kurtosis 4.02). The Bayes factor in favor of a mixture of two normals is about 20. The latter example is a dichotomous choice model of women's labor force participation ($n = 1555$) with conventional covariates. Of a dozen models including the conventional probit model and mixtures of up to five normals, the model with the highest marginal likelihood is a mixture of four normals. All the mixture models are highly favored relative to the conventional probit model, the Bayes factors ranging from 2×10^5 to 9×10^7.

6. Multivariate latent linear models

The natural extension of the ULLM to multiple decisions or outcomes is $\tilde{\boldsymbol{y}}_t = \boldsymbol{B}'\boldsymbol{x}_t + \boldsymbol{u}_t$, in which $\tilde{\boldsymbol{y}}_t$ has p elements and the $p \times 1$ disturbance vector $\boldsymbol{u}_t$ is i.i.d. and independent of the $\boldsymbol{x}_t$. Some (or all) of the elements of $\tilde{\boldsymbol{y}}_t$ may be fully observed, while others (or all) may be latent subject to known linear restrictions. The fully observed case is the seemingly unrelated regressions model [Zellner (1962)], the most widely applied multivariate econometric model. In many of these applications, such as neoclassical consumer and producer analysis, the observed outcome $\boldsymbol{y}_t = \tilde{\boldsymbol{y}}_t$ is subject to sum constraints that impose restrictions on $\boldsymbol{B}$ and render the distribution of $\boldsymbol{u}_t$ degenerate. In the multinomial probit model, the elements of $\tilde{\boldsymbol{y}}_t$ correspond to the unobserved utilities of p mutually exclusive choices. If choice j is made and observed, then $\tilde{y}_{jt} > \tilde{y}_{it}\ \forall\ i \neq j$, a set of $p-1$ linear restrictions. There must be an additional linear restriction for identification in this model as well, as detailed below in Section 6.1. In the standard selection model [Heckman (1979)] there are two equations ($p = 2$): in one equation the outcome $\tilde{y}_{1t}$ is fully observed if the latent variable $\tilde{y}_{2t} > 0$ in the other, whereas if $\tilde{y}_{2t} \leqslant 0$ then there is no information about $\tilde{y}_{1t}$. Extending the notation of Section 5, one either observes $y_{1t} = \tilde{y}_{1t}$ and $y_{2t} = (0, \infty)$, or $y_{1t} = (-\infty, \infty)$ and $y_{2t} = (-\infty, 0]$.

Unifying all of these models, and many more, is the multivariate latent linear model (MLLM) $\tilde{\boldsymbol{y}}_t = \boldsymbol{B}'\boldsymbol{x}_t + \boldsymbol{u}_t$. This model has several distinguishing characteristics. First, observed outcomes take the form $\boldsymbol{y}_t = \left[\tilde{\boldsymbol{y}}_t : \boldsymbol{c}_t^0 \leqslant \boldsymbol{F}_t^0 \tilde{\boldsymbol{y}}_t \leqslant \boldsymbol{d}_t^0\right]$. The vectors $\boldsymbol{c}_t^0$ and $\boldsymbol{d}_t^0$ and the $p \times p$ nonsingular matrices $\boldsymbol{F}_t^0$ are all known; elements of $\boldsymbol{c}_t^0$ and $\boldsymbol{d}_t^0$ are extended real numbers. Strict equalities are subsumed in this formulation. This includes not only the case of fully observed elements of $\tilde{\boldsymbol{y}}_t$, but also identities (for example, factor share equations) and identifying restrictions (for example, sum restrictions in the multinomial probit model). A second distinguishing characteristic of the model is that restrictions on the variance structure of $\boldsymbol{u}_t$ can be imposed. Some of these arise from identities among the elements of $\tilde{\boldsymbol{y}}_t$ that reduce the rank of $\mathrm{var}(\boldsymbol{u}_t)$ (e.g., share equations) while others are needed for identification (e.g., the selection and multinomial probit models). A third feature of the MLLM is that the disturbance distribution can be that of a continuous or finite mixture of multivariate normals. This extends the specification of the disturbance in the ULLM in a natural way.

As in the case of the ULLM, the treatment of the MLLM is entirely Bayesian with proper priors. This permits the full development of Bayes factors for comparing non-nested models with different assumptions about the distribution of the disturbances. Section 6.1 presents an overview of the MLLM, leaving technical detail to Appendix B. Section 6.2 provides some results with artificial data, to study the practicality of the methods.

6.1. An overview of the multivariate latent linear model

6.1.1. Linear restrictions in the multivariate latent linear model

In the MLLM the i.i.d. disturbances have a possibly degenerate scale mixture of normals distribution. The degeneracy arises if there is a $p \times g$ matrix $\boldsymbol{G}_0$ such that $\boldsymbol{G}_0'\tilde{\boldsymbol{y}}_t = \boldsymbol{g}_0\ \forall\ t$. For example, this may occur when $\tilde{\boldsymbol{y}}_t$ is an exhaustive set of input cost shares in production, or when $\tilde{\boldsymbol{y}}_t$ is an unobserved vector of utilities for all possible discrete choices. The potential degeneracy of the unconditional distribution for $\tilde{\boldsymbol{y}}_t$ has a number of implications for the MLLM. For the disturbances, $\boldsymbol{G}_0'\varepsilon_t \equiv \mathbf{0}\ \forall\ t$. Take $q = p - g$ and let $\boldsymbol{G}$ be any $p \times q$ orthonormal matrix of rank q such that $\boldsymbol{G}_0'\boldsymbol{G} = \mathbf{0}$. For example, when $\boldsymbol{G}_0' = (1,1)$ then $\boldsymbol{G}' = (1/\sqrt{2}, -1/\sqrt{2})$ or $\boldsymbol{G}' = (-1/\sqrt{2}, 1/\sqrt{2})$. The nondegenerate components of ε_t are $\varepsilon_t^* = \boldsymbol{G}'\varepsilon_t$, and the conditionally nondegenerate components of $\tilde{\boldsymbol{y}}_t$ are $\tilde{\boldsymbol{y}}_t^* = \boldsymbol{G}'\tilde{\boldsymbol{y}}_t$.

The restrictions $\boldsymbol{G}_0'\tilde{\boldsymbol{y}}_t = \boldsymbol{g}_0$ that reduce the rank of $\operatorname{var}(\boldsymbol{u}_t)$ also have implications for the row of $\boldsymbol{B}$ corresponding to the intercept term in $\boldsymbol{x}_t$. Linear restrictions on $\boldsymbol{B}$ can arise from other sources as well, however. For example in the multinomial probit model the need for additional, identifying restrictions on $\boldsymbol{B}$ arise if some covariates are specific to individuals. Additional restrictions may be implied by the underlying theory in any MLLM. Given ℓ_1 consistent and non-redundant linear restrictions on the pk elements of $\boldsymbol{B}$, there remain $\ell_2 = pk - \ell_1$, free parameters in $\boldsymbol{B}$.

With these restrictions in mind, rewrite the MLLM in the form

$$\tilde{\boldsymbol{y}}_t^* = \boldsymbol{B}^{*\prime}\boldsymbol{x}_t + \boldsymbol{u}_t^*, \tag{6.1}$$

in which $\tilde{\boldsymbol{y}}_t^* = \boldsymbol{G}'\tilde{\boldsymbol{y}}_t$, $\boldsymbol{B}^* = \boldsymbol{B}\boldsymbol{G}$, and $\boldsymbol{u}_t^* = \boldsymbol{G}'\boldsymbol{u}_t$.

6.1.2. Distribution of disturbances

In the MLLM the disturbance vectors $\boldsymbol{u}_t^*$ $(t = 1, \ldots, T)$ are i.i.d conditional on $\boldsymbol{x}_t$ $(t = 1, \ldots, T)$. The model subsumes several alternative assumptions about the distribution of $\boldsymbol{u}_t^*$, and we take up three in detail. The first is $\boldsymbol{u}_t^* \sim N(\mathbf{0}, \Sigma)$, in which Σ may be entirely free, or subject to a number of restrictions discussed in detail below and in Appendix B.

The second alternative assumption about the disturbance vector is $\boldsymbol{u}_t^* \sim t(\mathbf{0}, \Sigma; \lambda)$, the multivariate Student-t distribution with location vector $\mathbf{0}$, scale matrix Σ, and degrees of freedom parameter λ. The matrix Σ may again be free or subject to restrictions. The disturbances under this assumption may be represented $\boldsymbol{u}_t^* = \eta_t\sigma_{(t)}$, with $\left(\sigma_{(1)}, \ldots, \sigma_{(T)}\right)$ and the $p \times 1$ vectors $\eta_1, \ldots, \eta_T$ each i.i.d. and mutually independent conditional on $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T)$. The parameters $\sigma_{(t)}^2$ have the same inverted gamma distribution as in the ULLM, $\lambda/\sigma_{(t)}^2 \sim \chi^2(\lambda)$; $\eta_t \sim N(\mathbf{0}, \Sigma)$[32].

The third alternative assumption about the distribution of $\boldsymbol{u}_t^*$ is $\boldsymbol{u}_t^* \sim N\left(\alpha_j^*, \sigma_{j*}^2\Sigma\right)$ with probability p_j $(j = 1, \ldots, m)$; $\sum_{j=1}^m p_j = 1$. This multivariate normal mixture model is an extension of the one used in the ULLM – in fact, the marginal distribution of any linear combination $\boldsymbol{a}'\boldsymbol{u}_t^*$ has a univariate normal mixture distribution. The specification that the variance matrices in the mixture are all proportional to Σ is a restriction on the full multivariate normal mixture distribution, which would assign positive definitive matrices $\Sigma_1, \ldots, \Sigma_m$ as the variances of the respective states.

[32] For this construction see Johnson and Kotz (1972, Section 37.2).

We adopt the more restrictive assumption here because it allows some important simplifications in the MCMC algorithm constructed subsequently, and because our experience with the computations suggests that the full mixture model is poorly estimated in latent variable models with samples of the type and size typically available in economics.

In the normal mixture model the disturbances may be represented

$$\boldsymbol{u}_t^* = \boldsymbol{A}^{*\prime}\tilde{\boldsymbol{z}}_t + \sigma_{(t)}\eta_t.$$

The random vectors $\eta_1, \ldots, \eta_T$ are i.i.d. conditional on $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T)$: $\eta_t \sim N(\boldsymbol{0}, \Sigma)$. The latent random vectors $\left(\tilde{\boldsymbol{z}}_t^{\prime}, \sigma_{(t)}\right)$ are i.i.d. conditional on $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T)$ and $(\eta_1, \ldots, \eta_T)$. Their values are governed by a latent state variable $s(t)$. The joint distribution of $s(t)$, $\tilde{\boldsymbol{z}}_t$ and $\sigma_{(t)}$ is exactly the same as in the ULLM. Permutation of states is again prevented by means of the inequalities $\sigma_1 > \cdots > \sigma_m$.

All three specifications of the distribution of $\boldsymbol{u}_t^*$ in Equation (6.1) are embedded in

$$\underset{p\times 1}{\tilde{\boldsymbol{y}}_t^*} = \boldsymbol{A}^{*\prime} \underset{m\times 1}{\tilde{\boldsymbol{z}}_t} + \boldsymbol{B}^{*\prime} \underset{k\times 1}{\boldsymbol{x}_t} + \underset{p\times 1}{\varepsilon_t^*}, \qquad \varepsilon_t^* = \eta_t\sigma_{(t)},$$

in which $\eta_t \sim N(\boldsymbol{0}, \Sigma)$ is i.i.d. and independent of $\left(\tilde{\boldsymbol{z}}_t^{\prime}, \sigma_{(t)}\right)$, conditional on $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T)$. The normal, Student-$t$, and normal mixture models are distinguished by the distribution of $\sigma_{(t)}$. Only when $\boldsymbol{u}_t^*$ has a normal mixture distribution is $m > 0$.

6.1.3. Observed outcomes

As in the ULLM, corresponding to the latent outcomes $\tilde{\boldsymbol{y}}_t$ there is an observed, set-valued outcome $\boldsymbol{y}_t$ such that $\tilde{\boldsymbol{y}}_t \in \boldsymbol{y}_t$. The outcome $\boldsymbol{y}_t$ is determined solely by $\tilde{\boldsymbol{y}}_t$ so that

$$p(\boldsymbol{y}_t, \tilde{\boldsymbol{y}}_t|\boldsymbol{x}_t) = p(\tilde{\boldsymbol{y}}_t|\boldsymbol{x}_t)\, p(\boldsymbol{y}_t|\tilde{\boldsymbol{y}}_t) = p(\tilde{\boldsymbol{y}}_t|\boldsymbol{x}_t)\, \chi_{\boldsymbol{y}_t}(\tilde{\boldsymbol{y}}_t).$$

In the MLLM we take the form of $\boldsymbol{y}_t$ to be

$$\boldsymbol{c}_t \leqslant \boldsymbol{F}_t\tilde{\boldsymbol{y}}_t \leqslant \boldsymbol{d}_t \quad (t = 1, \ldots, T), \tag{6.2}$$

in which $\boldsymbol{c}_t$ and $\boldsymbol{d}_t$ are $q \times 1$ vectors of extended real numbers with $\boldsymbol{c}_t \leqslant \boldsymbol{d}_t$, and $\boldsymbol{F}_t$ is a $q \times p$ matrix of rank q.

This assumption subsumes quite a few models. The conventional seemingly unrelated regressions model (possibly with sum restrictions over equations, by means of $\boldsymbol{G}_0^{\prime}\tilde{\boldsymbol{y}}_t = \boldsymbol{g}_0$) is indicated by means of $\boldsymbol{c}_t = -\infty$, $\boldsymbol{d}_t = +\infty$, $\boldsymbol{F}_t = \left[\boldsymbol{I}_q : \boldsymbol{0}\right]$. The multinomial probit model with p exhaustive discrete choices has $q = p - 1$. If choice j is observed at t then $\boldsymbol{c}_t = \boldsymbol{0}$, $\boldsymbol{d}_t = +\infty$; each row i of $\boldsymbol{F}_t$ has two nonzero entries: $+1$ in column j, and -1 in column i if $i < j$ but in column $i + 1$ if $i \geqslant j$. In addition, $\boldsymbol{G}_0^{\prime} = (1, \ldots, 1)$ and $\boldsymbol{g}_0 = 0$. The sample selection model introduced by Heckman

(1979) has $p = q = 2$ and $\boldsymbol{F}_t = \boldsymbol{I}_2\ \forall\ t$. With the outcome equation first, $c_{1t} = -\infty$, $d_{1t} = +\infty$, $c_{2t} = -\infty$, $d_{2t} = 0$ if the outcome is unobserved; whereas $c_{1t} = d_{1t} = \tilde{y}_t$, $c_{2t} = 0$, $d_{2t} = +\infty$ if the outcome is observed. Many other models are possible. For example, a selection model with a dichotomous choice outcome has the same setup as the last example except that the first equation has either $c_{1t} = -\infty, d_{1t} = 0$ or $c_{1t} = 0, d_{1t} = +\infty$ if $c_{2t} = 0$ and $d_{2t} = +\infty$.

Augmenting Equation (6.2) with the constraints $\boldsymbol{G}_0'\tilde{\boldsymbol{y}}_t = \boldsymbol{g}_0$,

$$\begin{pmatrix} \boldsymbol{c}_t \\ \boldsymbol{g}_0 \end{pmatrix} \leqslant \begin{bmatrix} \boldsymbol{F}_t \\ \boldsymbol{G}_0' \end{bmatrix} \tilde{\boldsymbol{y}}_t \leqslant \begin{pmatrix} \boldsymbol{d}_t \\ \boldsymbol{g}_0 \end{pmatrix}.$$

Since

$$\begin{bmatrix} \boldsymbol{F}_t \\ \boldsymbol{G}_0' \end{bmatrix} \tilde{\boldsymbol{y}}_t = \begin{bmatrix} \boldsymbol{F}_t \\ \boldsymbol{G}_0' \end{bmatrix} \left[\boldsymbol{G} \vdots \boldsymbol{G}_0 \left(\boldsymbol{G}_0'\boldsymbol{G}_0\right)^{-1}\right] \begin{bmatrix} \boldsymbol{G}' \\ \boldsymbol{G}_0' \end{bmatrix} \tilde{\boldsymbol{y}}_t = \begin{bmatrix} \boldsymbol{F}_t\boldsymbol{G}\boldsymbol{G}'\tilde{\boldsymbol{y}}_t + \boldsymbol{F}_t\boldsymbol{G}_0\left(\boldsymbol{G}_0'\boldsymbol{G}_0\right)^{-1}\boldsymbol{g}_0 \\ \boldsymbol{g}_0 \end{bmatrix},$$

the constraints on the conditionally nondegenerate vector $\tilde{\boldsymbol{y}}_t^* = \boldsymbol{G}'\tilde{\boldsymbol{y}}_t$ are

$$\boldsymbol{c}_t - \boldsymbol{F}_t\boldsymbol{G}_0\left(\boldsymbol{G}_0'\boldsymbol{G}_0\right)^{-1}\boldsymbol{g}_0 \leqslant \boldsymbol{F}_t\boldsymbol{G}\tilde{\boldsymbol{y}}_t^* \leqslant \boldsymbol{d}_t - \boldsymbol{F}_t\boldsymbol{G}_0\left(\boldsymbol{G}_0'\boldsymbol{G}_0\right)^{-1}\boldsymbol{g}_0.$$

6.1.4. Prior distributions

Let β_2^* contain the free parameters in $\boldsymbol{B}$ remaining after imposition of ℓ_1 linear constraints. The benchmark prior distribution for β_2 is $\beta_2^* \sim N\left(\underline{\beta}_2^*, \underline{\boldsymbol{H}}_{\beta_2^*}^{-1}\right)$. Since this may not be a convenient or natural representation of prior information, Appendix B derives $\underline{\beta}_2^*$ and $\underline{\boldsymbol{H}}_{\beta_2^*}$, beginning from ℓ_1 linear restrictions, and independent normal prior distributions on at least $pk - \ell_1$ linear combinations of $\beta = \mathrm{vec}(\boldsymbol{B})$.

The prior distribution for Σ must account for the fact that in some important variants of the MLLM, scale restrictions on equations are necessary for identification. The best known example is the multinomial probit model, in which $\tilde{\boldsymbol{y}}_t^*$ is a vector of latent utilities. Scaling the utilities by a common positive factor produces no changes in $\boldsymbol{y}_t$, and so some convention is required to resolve the corresponding ambiguity in $\boldsymbol{A}$, $\boldsymbol{B}$, and Σ. So as not to require prior information that a certain coefficient is non-zero, and so as to maintain symmetry across equations, we normalize on the trace of the corresponding rows and columns of Σ. At the same time, we wish to cope with similar situations: for example, a multinomial probit model together with a selection equation; or, more generally, any situation in which the MLLM includes one or more sets of exhaustive discrete choices.

Including all of these cases requires some structure on $\boldsymbol{G}$ and Σ. Suppose that equations $i_1, \ldots, i_n$ in Equation (6.1) represent an exhaustive set of discrete choices. (There may be more than one such set.) Then there is a column of $\boldsymbol{G}_0$ with jth entry $\sum_{s=1}^{n} \delta_{j,i_s}$ $(j = 1, \ldots, p)$. Corresponding to this column of $\boldsymbol{G}_0$, choose $\boldsymbol{G}$ to have a subset of $n - 1$ columns, say $j_1, \ldots, j_{n-1}$, in which the only non-zero entries occur in those rows for which the corresponding columns of $\boldsymbol{G}_0$ have non-zero entries. All other columns of $\boldsymbol{G}$ can be chosen so that these rows have entries zero. For example, if $p = 6$, equations 1, 2, and 3 of the MLLM correspond to one set of exhaustive choices, and rows 5 and 6 correspond to another set, then the values of $\boldsymbol{G}_0$ and $\boldsymbol{g}_0$, and one choice for $\boldsymbol{G}$, are

$$\boldsymbol{G}_0 = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}, \quad \boldsymbol{g}_0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \boldsymbol{G} = \begin{bmatrix} 2/\sqrt{6} & 0 & 0 & 0 \\ -1/\sqrt{6} & 1/\sqrt{2} & 0 & 0 \\ -1/\sqrt{6} & -1/\sqrt{2} & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1/\sqrt{2} \\ 0 & 0 & 0 & -1/\sqrt{2} \end{bmatrix}.$$

Note that when $\boldsymbol{G}$ is chosen this way, entries $j_1, \ldots, j_{n-1}$ of ε_t^* are the nondegenerate disturbances associated uniquely with the set of exhaustive discrete choices in each case. Because $\boldsymbol{G}$ is orthonormal, $\sum_{s=1}^{n-1} \operatorname{var}\left(\varepsilon_{t,j_s}^*\right)$ is invariant to the particular choice of $\boldsymbol{G}$ within the constraints just described.

To indicate these entries of ε_t^* explicitly, construct the $q \times q$ matrix $\boldsymbol{U}$ as follows: $u_{ij} = 0 \; \forall j$ if entry i of ε_t^* does not correspond to any exhaustive set of discrete choices; $u_{ij} = 1$ if entries i and j of ε_t^* correspond to the same exhaustive set of discrete choices; and $u_{ij} = 0$ if entry i corresponds to an exhaustive set of discrete choices to which j does not correspond. Let $\boldsymbol{v}$ be a $q \times 1$ vector with entry $v_i = 0$ if entry i of ε_t^* corresponds to an exhaustive set of discrete choices and $v_i = 1$ otherwise.

We enforce a scaling constraint on each group of exhaustive discrete choices by requiring the sums of the variances of the corresponding entries of ε_t^* to be equal to one less than the number of choices in the group. We accomplish this by setting

$$\Sigma = \Delta(\Sigma^*) \cdot \Sigma^* \cdot \Delta(\Sigma^*).$$

The $q \times q$ matrix Σ^* is positive definite with typical element $\left[\sigma_{ij}^*\right]$. The $q \times q$ matrix $\Delta(\Sigma^*)$ is diagonal, with ith entry

$$\left(\frac{v_i + \sum_{j=1}^{q} u_{ij}}{v_i + \sum_{j=1}^{q} u_{ij}\sigma_{jj}^*}\right)^{1/2}.$$

Thus, if entries $j_1, \ldots, j_{n-1}$ of ε_t^* correspond to a set of exhaustive discrete choices, then $\sum_{s=1}^{n-1} \sigma_{j_s j_s} = n - 1$; and if entry j of ε_t^* does not correspond to such a group, $\sigma_{jj} = \sigma_{jj}^*$.

Finally, we employ a conventional inverted Wishart prior distribution, $\Sigma^* \sim W(\underline{\boldsymbol{S}}, \underline{\nu})$, for Σ^*. Appendix B shows that if $\underline{\boldsymbol{S}} = \underline{s}^2 \boldsymbol{I}_q$, then when $q < p$ the implied prior distribution for Σ is invariant to the particular choice of the orthonormal matrix $\boldsymbol{G}$.

This completes the prior distribution when the disturbances are normal, because in that case $\boldsymbol{B}$ and Σ are the only parameters in the model. The Student-t MLLM has one additional parameter, λ. The prior for λ is the same exponential distribution used in the ULLM, $\lambda \sim \exp(\underline{\lambda})$. The normal mixture model has three additional parameters: the vector of state probabilities $\boldsymbol{p}$, the state variances σ_j^2 $(j = 1, \ldots, m)$, and the matrix of state mean vectors $\boldsymbol{A}$. The first two of these play the same role in the MLLM and in the ULLM, and their prior distributions are of the same form. For $\boldsymbol{p}$ the distribution is Dirichlet: $p(\boldsymbol{p}) \propto \prod_{j=1}^{m} p_j^{r_j - 1}$. For the σ_j^2 the prior distributions are independent inverted gamma, $\underline{s}_j^2/\sigma_j^2 \sim \chi^2\left(\underline{\nu}_j\right)$, subject to the ordering restrictions $\sigma_1 > \cdots > \sigma_m$. For the same reasons discussed in the case of the ULLM it is productive to take the prior distribution of $\alpha^* = \operatorname{vec}(\boldsymbol{A}^*)$ to be Gaussian with mean $\mathbf{0}$, and variance proportional to $\Sigma : \alpha^* \sim N\left(\mathbf{0}, \Sigma \otimes \underline{\boldsymbol{H}}_{\alpha^*}^{-1}\right)$.

6.1.5. *Existence of the posterior distribution*

The existence of the posterior distribution in the MLLM follows in the same way that it does in the ULLM. In the normal and Student-t models, the data density is bounded. In the normal mixture model the data density is unbounded as a function of $\boldsymbol{A}$ and σ_j^2 $(j = 1, \ldots, m)$, but the existence of the posterior distribution can be demonstrated as in Section 5 and Appendix A, with minor variations. Similarly, as detailed in Appendix A, posterior moments exist if the corresponding prior moments exist, after infinitesimal reduction in the hyperparameters of the prior distributions of σ_j^2 $(j = 1, \ldots, m)$.

6.1.6. *MCMC algorithm for inference*

The explicit development of the posterior density kernel is presented in Appendix B. Due to the analytical intractability of the entire kernel, and the simple form of most of the conditional kernels, the Gibbs sampling algorithm is attractive here just as it is in the ULLM. There are eight groups of parameters or latent variables to which the algorithm is applied: $(\boldsymbol{A}^*, \boldsymbol{B}^*)$; Σ^*; $\sigma_{(t)}^2$ $(t = 1, \ldots, T)$; λ; $s(t)$ $(t = 1, \ldots, T)$ and $\tilde{\boldsymbol{Z}}$; $\boldsymbol{p}$; σ_j^2 $(j = 1, \ldots, m)$ and $\tilde{\boldsymbol{y}}_t$ $(t = 1, \ldots, T)$. As in the ULLM, not all parameters appear under each distributional assumption.

With one exception, the conditional posterior distributions in the MLLM come from the same families as their ULLM counterparts, although the parameters of these distributions are somewhat more involved as detailed in Appendix B. The conditional distribution of $(\boldsymbol{A}^*, \boldsymbol{B}^*)$ is multivariate normal; those of $\sigma_{(t)}^2$ $(t = 1, \ldots, T)$ are either trivial (normal and mixed normal) or are inverted gamma (Student-t). The $\tilde{\boldsymbol{y}}_t^*$ are conditionally independent truncated normal, so the algorithm in Geweke (1991) can be applied. In the Student-t model the functional form of the conditional posterior kernel for λ is the same as in the ULLM. In the normal mixture model the conditional distribution of $\boldsymbol{p}$ is Dirichlet, state assignments are multinomial, and the

σ_j^2 are respectively inverted gamma subject to the truncation restrictions imposed by $\sigma_1 > \cdots > \sigma_m$.

The exceptional conditional posterior distribution is that of Σ^* whenever there is degeneracy in $\tilde{\boldsymbol{y}}_t^*$ (i.e., $q < p$). This distribution is far from being inverted Wishart, and is of a non-standard form. We employ a Metropolis within Gibbs step, based on a tailored Gaussian approximation of the conditional posterior kernel. Details are furnished in Appendix B.

6.1.7. Marginal likelihoods

The Gelfand–Dey algorithm is used to approximate the marginal likelihood, just as in the ULLM. Once again, care is required in the handling of latent variables. In particular, it is essential to integrate across $\tilde{\boldsymbol{y}}_t^*$. This is trivial in the ULLM, but in the MLLM the computations, which are based on the GHK algorithm described in Section 2.1, are computationally more demanding. Details are given in Appendix B.

6.2. Some evidence from artificial data

Experiments with artificial data can provide some indication of how much information about the population is conveyed in the posterior distribution given a sample design. They can also provide a guide to the efficiency of the MCMC algorithm set forth in Section 6.1. Here we report the outcomes of experiments involving five variants of the MLLM. The first two involve no latent variables and serve as benchmarks for the other three. They are a three-equation multivariate regression model (Table 6.1) and a three-equation multivariate regression model in which the outcome variables always sum to unity (Table 6.2). The third variant of the MLLM is a multiple choice model (a multinomial probit model, when disturbances are normal) with three choices and covariates to mimic income and prices (Table 6.3). The fourth variant is a selection model, consisting of a dichotomous choice selection equation and an outcome equation with a continuously distributed dependent variable (Table 6.4). The final variant is also a selection model, but with a dichotomous dependent variable in the outcome equation (Table 6.5).

6.2.1. Multivariate linear model

Given the results for the ULLM in Table 5.1, there are few surprises in Table 6.1. Posterior standard deviations for coefficients consistently reflect the fact that $T = 2000$ in Table 5.1 and $T = 1000$ in Table 6.1. The same is true for the elements of Σ (the σ_{ij}) once account is taken of the fact that the posterior standard deviation of σ^2 in Table 5.1 is approximately twice that of σ. The most notable contrast in comparing the MLLM fully observed outcome model with its ULLM counterpart is in the greatly increased computational efficiencies for p_1 and σ_2 in the normal mixture model. This reflects the fact that classification by state is substantially more certain for a vector of random

Table 6.1
Multivariate full rank linear model ($T = 1000$)

A. Model specification

$$\boldsymbol{y}_t = \begin{pmatrix} y_{1t}\\ y_{2t}\\ y_{3t}\end{pmatrix} = \begin{bmatrix}\beta_{11} & \beta_{21}\\ \beta_{12} & \beta_{22}\\ \beta_{13} & \beta_{23}\end{bmatrix}\begin{pmatrix}1\\ x_{2t}\end{pmatrix} + \begin{pmatrix}u_{1t}\\ u_{2t}\\ u_{3t}\end{pmatrix} = \boldsymbol{B}'\boldsymbol{x}_t + \eta_t\sigma_{(t)}; \quad \eta_t \overset{\text{i.i.d.}}{\sim} N(\mathbf{0},\Sigma); \Sigma = \left[\sigma_{ij}\right] = \boldsymbol{I}_3$$

Population for all variants: $x_{2t} \overset{\text{i.i.d.}}{\sim} N(0,1)$, $\beta_{ij} = 1$ ($j = 1,2,3$; $i = 1,2$)

Normal disturbances: $\sigma_{(t)} \equiv 1$

Student-t disturbances: $4/\sigma^2_{(t)} \overset{\text{i.i.d.}}{\sim} \chi^2(4)$

Mixed normal disturbances: $\sigma^2_{(t)} = 1(p = .6)$ or $\sigma^2_{(t)} = .04(p = .4)$

B. Prior distributions and moments

Parameters	Prior distribution	Prior mean	Prior s.d.
β_{ij} ($j = 1,2,3$; $i = 1,2$)	$\beta_{ij} \sim N(0,25)$	0.0	5.0
$\Sigma = \Sigma^*$	$\Sigma^* \sim IW\,(5\boldsymbol{I}_3, 5)$	$5\boldsymbol{I}_3$	∞
λ	Exponential	4	2.828
α	$\alpha^* \sim N\,(\mathbf{0}, \Sigma \otimes 5\boldsymbol{I}_2)$	0	$2.236\,(\sigma_{jj})^{1/2}$
σ_j^2	$5/\sigma_1^2 \sim \chi^2(5)$	1.189	1.291
	$.2/\sigma_2^2 \sim \chi^2(5)$	0.238	0.258
p_1	Beta(5,5)	0.500	0.204

Some posterior moments

Parameter	Mean	Stan. dev.	RNE	Parameter	Mean	Stan. dev.	RNE
C. Normal disturbances; .384 sec./iter.							
$\beta_{22} = 1$	0.951	0.032	0.876	$\sigma_{11} = 1.0$	1.020	0.046	0.744
				$\sigma_{12} = 0.0$	0.043	0.033	1.811
D. Student-t disturbances; .400 sec./iter.							
$\beta_{22} = 1$	0.945	0.031	0.828	$\sigma_{11} = 1.0$	1.029	0.061	0.237
$\lambda = 4$	4.392	0.361	0.095	$\sigma_{12} = 0.0$	0.043	0.033	0.581
E. Normal mixture disturbances; .416 sec./iter.							
$\beta_{22} = 1$	1.010	0.009	0.329	$\sigma_1^2 \cdot \sigma_{11} = 1.0$	1.029	0.061	0.237
$\beta_{13} + \alpha_1 = 3$	3.003	0.043	1.177	$\sigma_1^2 \cdot \sigma_{21} = 0.0$	0.043	0.033	0.581
$\beta_{13} + \alpha_2 = 1.4$	1.383	0.009	0.763	$\sigma_2^2 \cdot \sigma_{11} = .04$	0.038	0.002	1.221
$p_1 = .6$	0.577	0.015	1.685	$\sigma_2^2 \cdot \sigma_{21} = 0$	–.002	0.001	2.989

Table 6.2
Degenerate multivariate linear model (T = 1000)

A. Model specification

$$\boldsymbol{y}_t = \boldsymbol{B}'\boldsymbol{x}_t + \tilde{\eta}_t\sigma_{(t)}; \quad \eta_t \overset{\text{i.i.d.}}{\sim} N(\boldsymbol{0}, \Psi); \quad \Psi = \left[\psi_{ij}\right]$$

Population for all variants: $x_{1t} = 1$; $x_{2t} \overset{\text{i.i.d.}}{\sim} N(0,1)$; $\boldsymbol{B} = \begin{bmatrix} .3 & .2 & .5 \\ 1 & -.5 & -.5 \end{bmatrix}$; $\Psi = \begin{bmatrix} 1 & -.5 & -.5 \\ -.5 & .5 & 0 \\ -.5 & 0 & .5 \end{bmatrix}$

Normal disturbances: $\sigma_{(t)} \equiv 1$

Student-*t* disturbances: $3.5/\sigma^2_{(t)} \overset{\text{i.i.d.}}{\sim} \chi^2(3.5)$

Mixed normal disturbances: $\sigma^2_{(t)} = 1$ ($p = .6$) or $\sigma^2_{(t)} = .02$ ($p = .4$)

B. Prior distributions and moments

Parameters	Prior distribution	Prior mean	Prior s.d.
β_{ij} ($j = 1,2,3; i = 1,2$)	$\beta_{ij} \sim N(0,25)$ $\sum_{j=1}^{3} \beta_{1j} = 1$, $\sum_{j=1}^{3} \beta_{2j} = 0$	0.0	5.0
$\Sigma = \Sigma^*$	$\Sigma^* \sim IW(5\boldsymbol{I}_2, 5)$	$2.5\boldsymbol{I}_2$	2.236 ($i = j$); 1.673 ($i \neq j$)
$\lambda, \alpha, \sigma_j, p_1$	As in Table 6.1		

Some posterior moments

Parameter	Mean	Stan. dev.	RNE	Parameter	Mean	Stan. dev.	RNE
C. Normal disturbances; .278 sec./iter.							
$\beta_{21} = 1.0$	0.989	0.016	1.248	$\psi_{11} = 1.0$	1.044	0.046	1.096
$\beta_{22} = -0.5$	–.501	0.012	1.731	$\psi_{21} = -0.5$	–.538	0.029	0.866
$\beta_{23} = -0.5$	–.488	0.011	0.810	$\psi_{31} = -0.5$	–.506	0.028	2.077
D. Student-t disturbances; .289 sec./iter.							
$\beta_{21} = 1.0$	0.981	0.019	0.524	$\psi_{11} = 1.0$	1.027	0.073	0.173
$\beta_{22} = -0.5$	–.487	0.014	1.356	$\psi_{21} = -0.5$	–.502	0.038	0.302
$\beta_{23} = -0.5$	–.494	0.014	0.511	$\psi_{31} = -0.5$	–.526	0.041	0.188
$\lambda = 3.5$	3.808	0.350	0.161				
E. Normal mixture disturbances; .309 sec./iter.							
$\beta_{21} = 1.0$	0.993	0.005	1.172	$\sigma_1^2 \cdot \psi_{11} = 1.0$	0.975	0.057	0.702
$\beta_{22} = -0.5$	–.494	0.004	1.356	$\sigma_1^2 \cdot \psi_{21} = -0.5$	–.492	0.038	0.312
$\beta_{23} = -0.5$	–.499	0.004	0.764	$\sigma_1^2 \cdot \psi_{31} = -0.5$	–.483	0.042	0.288
$\beta_{11} + \alpha_1 = 0.3$	0.239	0.042	1.522	$\sigma_2^2 \cdot \psi_{11} = .04$	0.039	0.003	0.457
$\beta_{11} + \alpha_2 = 0.3$	0.318	0.012	0.589	$\sigma_2^2 \cdot \psi_{21} = -.02$	–.020	0.002	0.652
$p_1 = .6$	0.569	0.020	0.489	$\sigma_2^2 \cdot \psi_{31} = -.02$	–.019	0.002	0.539

variables than for a scalar, given the same underlying mixture of normals population. For much the same reason, there is substantially more information about the degrees of freedom parameter λ in the Student-t distribution in the multivariate case than in the univariate. This is borne out in the reduced posterior standard deviation and increased RNE for λ in Table 6.1, panel D, as opposed to its counterpart in panel D of Table 5.1. It reflects substantially reduced posterior uncertainty for the latent variables $\sigma_{(t)}$.

6.2.2. Degenerate multivariate linear model

The second variant of the MLLM is the degenerate three-equation linear model described in Table 6.2. For each observation the dependent variables always sum to one, as they do in share equations for producer factor demand. In the notation of Section 6.1, $\boldsymbol{G}'_0 = (1, 1, 1)$ and $\boldsymbol{g}_0 = 1$. The prior distribution imposes the corresponding restrictions on the intercepts, $\beta_{11} + \beta_{12} + \beta_{13} = 1$, and the covariate coefficients, $\beta_{21} + \beta_{22} + \beta_{23} = 0$. The posterior means for the covariate coefficients shown in panels C, D, and E of Table 6.2 of course reflect these restrictions. The coefficient posterior standard deviations are lower than for the coefficients in Table 6.1. This reflects both the reduction in disturbance variance ($tr(\Sigma) = 3$ in Table 6.1 whereas $tr(\Psi) = 2$ in Table 6.2) and also the additional information provided by the coefficient sum restrictions. Posterior moments for the first column of $\Psi = \text{var}(\varepsilon_t)$ are provided in Table 6.2. Posterior means sum to zero, reflecting the degeneracy in $\boldsymbol{y}_t$. The posterior standard deviation of ψ_{11} in Table 6.2 is about the same as that of $\sigma_{11} = 1$ in Table 6.1, reflecting the fact that $\psi_{11} = \sigma_{22} + \sigma_{33} = 1$. That for ψ_{21} is smaller than for σ_{21}, because $\psi_{11} \cdot \psi_{22} = \frac{1}{2}$ whereas $\sigma_{11} \cdot \sigma_{22} = 1$ in Table 6.1.

6.2.3. Multiple discrete choice model

The third variant of the MLLM considered here is a three-choice linear model. In this model, the latent $\tilde{\boldsymbol{y}}_t$ can be interpreted as a vector of utilities for each of p discrete choices. The sample size is $T = 2000$. Panel A of Table 6.3A provides the model specification. The sample design implies equal unconditional choice probabilities. The covariate x_{t2} mimics an income variable: observations with high x_{2t} tend to choose 3, while low values tend to make choice 1 or 2. The covariates x_{t3}, x_{t4} and x_{t5} mimic prices, with choice 2 being more price responsive than choices 1 or 3. Overall, choices are strongly driven by the covariates, with the disturbance providing 20% of the variance in the utility differential between choices 1 and 2, 7.7% for choices 1 and 3, and 3.3% for choices 2 and 3.

The prior distribution incorporates identifying restrictions, using the approach developed in Section 6.1. Coefficients on the common covariates (the intercept and x_{2t}) and the disturbances (ε_{jt}) must sum to zero, or equivalently $\tilde{y}_{1t} + \tilde{y}_{2t} + \tilde{y}_{3t} = \beta_{31}x_{t3} + \beta_{42}x_{t4} + \beta_{53}x_{t5}$. This prevents systematic addition of multiples of any covariate(s) to any equations(s) without changing the probabilities of the choices. The 2×1 nondegenerate shock ε_t^* with variance Σ is subject to the restriction that $\text{tr}(\Sigma) = 2$, using the transformation presented in Section 6.1[33]. Notice that these identifying restrictions are symmetric across all three equations, as are the prior distributions set out in panel B of Table 6.3A.

In the MLLM, it is important to monitor the convergence of the MCMC algorithm when there is substantial latency in the outcome variables, as is the case here. As currently implemented, the MCMC algorithm takes as initial values prior means of parameters. About 9000 iterations are required for convergence of all parameters, as assessed informally by examining the simulations. The posterior moments in panels C, D and E of Table 6.3B are computed from the last 10 000 iterations out of 22 000. The RNEs reported are based on every tenth iteration; RNEs based on every iteration would be lower, and direct comparisons of the RNEs in Table 6.3B with those in Tables 6.1 or 6.2 cannot be made.

As must be the case, posterior means reflect the identifying restrictions on the β_{ij}. The identifying adding up restriction leads to posterior means of column sums of $\Psi = \mathrm{var}(\varepsilon_t)$ being zero, and the posterior mean of $\mathrm{tr}(\Psi) = \mathrm{tr}(\Sigma)$ being identically two. The substantial information about covariate coefficients in the sample design is reflected in posterior standard deviations that are one-tenth or less of posterior mean in most cases. All posterior means in Table 6.3B are within two posterior standard deviations of population values.

The posterior standard deviations of the covariate coefficients in this model are nearly an order of magnitude higher than in the linear models with fully observed outcomes (Tables 6.1 and 6.2). There is no such increase in the case of the ULLM (Table 5.3 versus Table 5.1). By contrast, the posterior standard deviations of the variances and covariances ψ_{ij} are about the same for the three choice models as they were in the linear models described in Tables 6.1 and 6.2. Of course, these standard deviations reflect uncertainty after the imposition of the trace and singularity constraints on Ψ, which leaves only two free parameters.

Comparison of panels C, D and E in Table 6.3B reveals that posterior standard deviations of common parameters are about the same, for the three different distributions of the disturbances. This observation, combined with the comparison with the linear models in Tables 6.1 and 6.2, strongly suggests that the latency of $\tilde{\boldsymbol{y}}_t$ is the dominant source of uncertainty about the parameters, and that additional uncertainty contributed by a more flexible distribution of the disturbances is minor by comparison. On the other hand, for the common parameters RNE is highest for normal disturbances (panel C) and lowest for mixed normal disturbances (panel E), without exception. This reflects the additional steps introduced into the Gibbs sampling algorithm for the Student-*t* and normal mixture models, increasing serial correlation in the draws of the parameters.

[33] In the current computer code that implements the MLLM, the transformations to deal with degeneracy and the variance trace restrictions can be made transparent to the user. The user need only indicate the restrictions on the coefficients, the fact that the disturbances sum to zero, and the fact that the trace of the variance matrix sums to two.

Table 6.3A
Multiple discrete choice model ($T = 2000$)

A. Model specification

$\boldsymbol{y}_t = \boldsymbol{B}'\boldsymbol{x}_t + \tilde{\eta}_t\sigma_{(t)}$; $\tilde{\eta}_t \overset{\text{i.i.d.}}{\sim} N(\boldsymbol{0}, \Psi)$; $\Psi = \left[\psi_{ij}\right]$; $\boldsymbol{y}_t = \sum_{j=1}^{3} \chi_{S_j}(\tilde{\boldsymbol{y}}_t) S_j$; $S_j = \left\{\tilde{\boldsymbol{y}}_t : \tilde{y}_{tj} \geqslant \tilde{y}_{ti}\ (i = 1,2,3)\right\}$

Population for all variants: $x_{1t} = 1$; $x_{2t} \overset{\text{i.i.d.}}{\sim} N(1,1)$, $x_{jt} \sim N(0,1)$ $(j = 3,4,5)$;

$$\boldsymbol{B}' = \begin{bmatrix} 1 & -1 & -1 & 0 & 0 \\ 2 & -2 & 0 & -2 & 0 \\ -3 & 3 & 0 & 0 & -1 \end{bmatrix}; \Psi = \begin{bmatrix} 1 & -.5 & -.5 \\ -.5 & .5 & 0 \\ -.5 & 0 & .5 \end{bmatrix}$$

Normal disturbances: as in Table 6.2

Student-t disturbances: as in Table 6.2

Mixed normal disturbances: as in Table 6.2

B. Prior distributions and moments

Parameters	Prior distribution	Prior mean	Prior s.d.
β_{ij} $(i \leqslant 2)$	$\beta_{ij} \sim N(0, 25)$	0.0	5.0
$\beta_{i+2,i}$	$\beta_{ij} \sim N(-1, 25)$	−1.0	5.0
β_{ij} $(i > 2$ and $i \neq j + 2)$	$\beta_{ij} = 0$	0.0	0.0
	$\sum_{j=1}^{3} \beta_{1j} = \sum_{j=1}^{3} \beta_{2j} = 0$		
σ_j	$\sigma_1 = 1$	1.0	0.0
	$.2/\sigma_2^2 \sim \chi^2(5)$	0.238	0.258
Σ^*	As in Table 6.2		
λ, α, p_1	As in Table 6.1		

6.2.4. Continuous selection model

The fourth variant taken up in these experiments is the conventional selection model due to Heckman (1979). Table 6.4 presents the model specification and posterior moments for a particular set of variants on this model. There are two equations. The first equation is an outcome equation with a continuously distributed dependent variable. (For example, this dependent variable could be a wage rate.) The second equation is a dichotomous choice equation. (In the example, this choice could be the decision to take a job.) The dependent variable in the outcome equation is observed or not depending on the value taken by the choice variable in the choice equation. (In the example, the wage rate is observed only if the job is taken.) If there is correlation between the disturbances in the two equations then conventional single equation methods for inference applied to the outcome equation, ignoring the sample selection process, are misleading. In the example presented in Table 6.4 each equation has an intercept but the covariates are different and independent.

The artificial data set is designed so that fewer than half of the outcomes are actually observed. This is achieved by means of the negative value of the intercept in the choice

Table 6.3B
Multiple discrete choice model (T = 2000), *continued*

Some posterior moments

Parameter	Mean	Stan. dev.	RNE	Parameter	Mean	Stan. dev.	RNE
C. Normal disturbances; .985 sec./iter.							
$\beta_{21} = -1.0$	−.883	0.112	0.153	$\beta_{31} = -1$	−1.048	0.069	1.170
$\beta_{22} = -2.0$	−2.097	0.111	0.252	$\beta_{42} = -2.0$	−2.126	0.108	0.243
$\beta_{23} = 3.0$	2.979	0.143	0.156	$\beta_{53} = -1.0$	−1.061	0.071	0.208
$\psi_{11} = 1.0$	0.965	0.051	0.508	$\psi_{21} = -0.5$	−.599	0.065	0.050
$\psi_{22} = 0.5$	0.634	0.072	0.068	$\psi_{31} = -0.5$	−.366	0.072	0.068
$\psi_{33} = 0.5$	0.401	0.065	0.050	$\psi_{32} = 0.0$	−.035	0.065	0.508
D. Student-t disturbances; 1.01 sec./iter.							
$\beta_{21} = -1.0$	−.922	0.125	0.062	$\beta_{31} = -1$	−1.005	0.097	0.095
$\beta_{22} = -2.0$	−2.077	0.177	0.064	$\beta_{42} = -2.0$	−2.110	0.182	0.048
$\beta_{23} = 3.0$	2.998	0.215	0.069	$\beta_{53} = -1.0$	−1.019	0.103	0.132
$\psi_{11} = 1.0$	1.031	0.048	0.343	$\psi_{21} = -0.5$	−.533	0.070	0.045
$\psi_{22} = 0.5$	0.502	0.077	0.056	$\psi_{31} = -0.5$	−.498	0.077	0.055
$\psi_{33} = 0.5$	0.467	0.070	0.045	$\psi_{32} = 0.0$	0.031	0.048	0.343
$\lambda = 3.5$	3.645	0.882	0.055				
E. Normal mixture disturbances; .784 sec./iter.							
$\beta_{21} = -1.0$	−1.001	0.127	0.024	$\beta_{31} = -1$	−1.068	0.087	0.013
$\beta_{22} = -2.0$	−2.007	0.163	0.009	$\beta_{42} = -2.0$	−2.054	0.157	0.009
$\beta_{23} = 3.0$	3.008	0.261	0.008	$\beta_{53} = -1.0$	−1.091	0.095	0.011
$\beta_{11} + \alpha_1 = 1.0$	1.091	0.154	0.011	$\beta_{12} + \alpha_1 = 2.0$	1.935	0.186	0.009
$\beta_{11} + \alpha_2 = 1.0$	0.944	0.186	0.009	$\beta_{12} + \alpha_2 = 2.0$	2.039	0.165	0.008
$\psi_{11} = 1.0$	0.993	0.060	0.147	$\psi_{21} = -0.5$	−.546	0.079	0.025
$\psi_{22} = 0.5$	0.553	0.078	0.039	$\psi_{31} = -0.5$	−.447	0.078	0.039
$\psi_{33} = 0.5$	0.454	0.079	0.025	$\psi_{32} = 0.0$	−.007	0.060	0.147
$\sigma_2 = 0.2$	0.208	0.057	0.122	$p_1 = 0.6$	0.624	0.072	0.014

equation, and in the artificial sample of size 1000, only 254 outcomes in the first equation are actually observed. The posterior standard deviations in the probit choice equation are comparable to those in the probit model (Table 6.4, panel C compared with Table 5.3, panel C), if one keeps in mind the difference in sample size (T = 1000 here, but T = 2000 in the ULLM experiments). In the Student-t model they are somewhat lower, and in the normal mixture model, somewhat higher. In the outcome equation

Table 6.4
Continuous outcome selection model ($T = 1000$)

A. Model specification

$$\tilde{\boldsymbol{y}}_t = \boldsymbol{B}'\boldsymbol{x}_t + \eta_t\sigma_{(t)}; \quad \eta_t \overset{\text{i.i.d.}}{\sim} N(\boldsymbol{0},\Sigma); \quad \boldsymbol{y}_t = \chi_{(0,\infty)}(\tilde{y}_{1t})\begin{pmatrix}\tilde{y}_{1t}\\(0,\infty)\end{pmatrix} + \chi_{(-\infty,0)}(\tilde{y}_{2t})\begin{pmatrix}(-\infty,\infty)\\(-\infty,0)\end{pmatrix}$$

Population for all variants: $x_{jt} \sim N(0,1)$ $(j = 2,3)$; $\boldsymbol{B}' = \begin{bmatrix}1 & 1 & 0\\-1 & 0 & 1\end{bmatrix}$; $\Sigma = \begin{bmatrix}1 & .5\\.5 & 1\end{bmatrix}$

Normal disturbances: $\sigma_{(t)} \equiv 1$

Student-*t* disturbances: $4/\sigma^2_{(t)} \overset{\text{i.i.d.}}{\sim} \chi^2(4)$

Mixed normal disturbances: $\sigma^2_{(t)} = 1$ $(p = .6)$ or $\sigma^2_{(t)} = .04$ $(p = .4)$

B. Prior distributions and moments

Parameters	Prior distribution	Prior mean	Prior s.d.
$\beta_{11}, \beta_{12}, \beta_{21}, \beta_{32}$	$\beta_{ij} \sim N(0, 25)$	0.0	5.0
β_{31}, β_{22}	$\beta_{ij} = 0$	0.0	0.0
Σ^*	$\Sigma^* \sim IW(5\boldsymbol{I}_2, 5)$	$2.5\boldsymbol{I}_2$	2.236 $(i = j)$, 1.673 $(i \neq j)$
λ, α, σ_j, p_1	As in Table 6.1		

Some posterior moments

Parameter	Mean	Stan. dev.	RNE	Parameter	Mean	Stan. dev.	RNE
C. Normal disturbances; .385 sec./iter.							
$\beta_{11} = 1.0$	0.973	0.114	0.012	$\beta_{12} = -1.0$	−.923	0.056	0.095
$\beta_{21} = 1.0$	0.977	0.055	0.038	$\beta_{32} = 1.0$	0.951	0.065	0.050
$\sigma_{11} = 1.0$	0.988	0.108	0.018	$\sigma_{21} = .5$	0.582	0.102	0.012
D. Student-t disturbances; .398 sec./iter.							
$\beta_{11} = 1.0$	1.289	0.124	0.016	$\beta_{12} = -1.0$	−1.099	0.089	0.026
$\beta_{21} = 1.0$	1.070	0.073	0.084	$\beta_{32} = 1.0$	1.209	0.103	0.023
$\sigma_{11} = 1.0$	0.889	0.154	0.026	$\sigma_{21} = .5$	0.241	0.104	0.017
$\lambda = 4.0$	3.29	0.618	0.013				
E. Normal mixture disturbances; .400 sec./iter.							
$\beta_{11} + \alpha_1 = 1.0$	0.907	0.076	0.031	$\beta_{21} = 1.0$	0.951	0.022	0.104
$\beta_{11} + \alpha_2 = 1.0$	1.035	0.024	0.115	$\beta_{32} = 1.0$	1.131	0.074	0.005
$\sigma_{11} = 1.0$	1.100	0.112	0.026	$\sigma_{21} = .5$	0.577	0.100	0.017
$\sigma_2 = 0.2$	0.203	0.021	0.032	$p_1 = 0.6$	0.625	0.038	0.018

coefficient posterior standard deviations are about triple what they were in Table 5.1, in all three models (panels C, D and E). This is somewhat higher than that accounted for by the difference between the sample size in Table 5.3 and the observed first equation outcomes in Table 6.4.

The posterior mean of the covariance between the disturbances in the choice and outcome equations is within two posterior standard deviations of the population value, when the disturbances are normal or mixed normal. In the Student-t model the posterior mean is 2.5 posterior standard deviations less than the population value, perhaps a reflection of the low posterior mean for λ. The uncertainty about this parameter, as well as the variance in the outcome equation, is about the same for all three variants of the disturbance specification. The posterior information about the degrees of freedom parameter λ in the selection model, is about the same as in the univariate linear model with double the number of observations (Table 5.1, panel D), and substantially greater than in the binary choice model (Table 5.3, panel D). The comparison with Table 5.1 indicates that the increased information about alternatives to normality in a vector (as opposed to a scalar, noted previously) carries over to situation in which the additional component is only partially observed. After accounting for sample sizes there is a similar comparison for the parameters of the mixed normal distribution (Panel E of Table 6.4 compared with its counterparts in Tables 5.1 and 5.3).

6.2.5. Discrete choice selection model

The final variant of the MLLM is similar to the conventional selection model, except that the outcome is a discrete binomial choice rather than a continuous random variable. For example, the second equation could model the decision to take a job, as in the previous example. The first equation could indicate the selection of public or private transit as the mode of commuting. This example could be extended to the case of multinomial choice conditional on selection – for example, transit choices could be divided more finely (train, bus, automobile, other private). Beyond the dichotomous outcome, the econometric structure of this example differs from the previous one in the inclusion of the first covariate in both equations. With two covariates, one in the outcome equation and both in the selection equation, this structure corresponds to the simplest selection model. Because the outcome equation is a dichotomous choice, the variance of each equation is normalized to 1.0. This provides the simplest example of restrictions on the traces of two subsets of the parameter matrix Σ^* described in Section 6.1.

Compared with the continuous selection model, there is a loss in information in the discrete choice selection model due to the fact that even when the outcome is selected, only the sign of $\tilde{y}_{t1}$ is observed. There is a gain in information because the outcome equation variance is normalized to one. On the whole, there appear to be no systematic differences in posterior standard deviations given normal or Student-t disturbances (panels C and D of Table 6.5 compared with their counterparts in Table 6.4). In the mixed normal model there is substantially increased uncertainty

Table 6.5
Dichotomous choice selection model (T = 2000)

A. Model specification

$$\tilde{\boldsymbol{y}}_t = \boldsymbol{B}'\boldsymbol{x}_t + \eta_t \sigma_{(t)}; \quad \eta_t \overset{\text{i.i.d.}}{\sim} N(\boldsymbol{0}, \Sigma);$$

$$\boldsymbol{y}_t = \chi_{(0,\infty)}(\tilde{y}_{1t}) \begin{pmatrix} \chi_{(0,\infty)}(\tilde{y}_{1t})\,(0,\infty) + \chi_{(-\infty,0)}(\tilde{y}_{1t})\,(-\infty,0) \\ (0,\infty) \end{pmatrix} + \chi_{(-\infty,0)}(\tilde{y}_{2t}) \begin{pmatrix} (-\infty,\infty) \\ (-\infty,0) \end{pmatrix}$$

Population for all variants: $x_{jt} \sim N(0,1)$ $(j = 2,3)$; $\boldsymbol{B}' = \begin{bmatrix} 1 & 1 & 0 \\ -1 & -0.8 & 1 \end{bmatrix}$; $\Sigma = \begin{bmatrix} 1 & .5 \\ .5 & 1 \end{bmatrix}$

Normal disturbances: $\sigma_{(t)} \equiv 1$

Student-t disturbances: $4/\sigma^2_{(t)} \overset{\text{i.i.d.}}{\sim} \chi^2(4)$

Mixed normal disturbances: $\sigma^2_{(t)} = 1$ $(p = .6)$ or $\sigma^2_{(t)} = .04$ $(p = .4)$

B. Prior distributions and moments

Parameters	Prior distribution	Prior mean	Prior s.d.
$\beta_{11}, \beta_{12}, \beta_{21}, \beta_{22}, \beta_{32}$	$\beta_{ij} \sim N(0,25)$	0.0	5.0
β_{22}	$\beta_{22} = 0$	0.0	0.0
Σ^*, λ, α, σ_j, p_1	As in Table 6.2		

Some posterior moments

Parameter	Mean	Stan. dev.	RNE	Parameter	Mean	Stan. dev.	RNE
C. Normal disturbances; .410 sec./iter.							
$\beta_{11} = 1.0$	0.951	0.084	0.067	$\beta_{12} = -1.0$	−1.013	0.066	0.395
$\beta_{21} = 1.0$	1.017	0.059	0.145	$\beta_{22} = -0.8$	−.837	0.044	0.605
$\sigma_{12} = 0.5$	0.535	0.039	0.008	$\beta_{32} = 1.0$	0.989	0.046	0.280
D. Student-t disturbances; .725 sec./iter.							
$\beta_{11} = 1.0$	1.240	0.145	0.038	$\beta_{12} = -1.0$	−1.098	0.113	0.199
$\beta_{21} = 1.0$	1.061	0.089	0.068	$\beta_{22} = -0.8$	−.831	0.080	0.147
$\sigma_{12} = 0.5$	0.260	0.129	0.028	$\beta_{32} = 1.0$	1.022	0.086	0.170
$\lambda = 4.0$	4.526	2.038	0.075				
E. Normal mixture disturbances; .400 sec./iter.							
$\beta_{11} + \alpha_1 = 1.0$	0.907	0.200	0.005	$\beta_{21} = 1.0$	0.928	0.148	0.004
$\beta_{11} + \alpha_2 = 1.0$	0.944	0.167	0.005	$\beta_{22} = -0.8$	−.797	0.100	0.004
$\sigma_{12} = 0.5$	0.509	0.131	0.011	$\beta_{32} = 1.0$	1.003	0.124	0.004
$\sigma_2 = 0.2$	0.207	0.045	0.022	$p_1 = 0.6$	0.550	0.064	0.005

about most parameters, and posterior standard deviations are greater in every case. This is consistent with the sharp increase noted between the degenerate multivariate linear model and the multiple discrete choice model, in the case of mixed normal disturbances, and can be attributed to the failure to ever fully observe any element of the outcome vector $\tilde{\boldsymbol{y}}_t$.

6.2.6. Marginal likelihood approximations

In the case of the MLLM, marginal likelihood approximations are trivial when $\tilde{\boldsymbol{y}}_t$ is fully observed, because the likelihood function can be evaluated in closed form. When this is the case, as it is in the first two of the five examples taken up here, the prior and data density can be evaluated every iteration with negligible increase in computing time. In general, when $\tilde{\boldsymbol{y}}_t$ has latent components, a simulation algorithm like the GHK algorithm described in Section 2.1 must be applied. The GHK algorithm, as well as variants on this algorithm and various alternatives to it, produce simulation consistent rather than unbiased evaluations of the log data density. Thus it is necessary to embed an iterative procedure within an iterative procedure, and computation time can become prohibitive if data density evaluations are made each iteration. Whether or not this is the case depends on the structure of the latency in $\tilde{\boldsymbol{y}}_t$. For example, in the selection model with a fully observed outcome, an essentially closed form evaluation of the data density is still possible. The extended GHK algorithm described in Appendix B exploits this fact and converges in a single iteration. At the other extreme, the three choice models with Student-*t* distributions have data density evaluations that are by far the most time consuming using the methods presented here: on average, a single evaluation required 15 minutes. For normal and normal mixture distributions average evaluation time was about four minutes. To make the procedure practical, data density evaluations were made every 100th iteration. For data density evaluation, the MCMC algorithm was executed for 22 000 iterations in the case of the multiple discrete choice model and single discrete choice outcome selection model, and 12 000 iterations in the case of the continuous outcome selection model. In every case, the last 10 000 iterations were used, so the results presented here are based on 100 evaluations of the prior and data densities. The loss in information due to using every 100th iteration (as opposed to each iteraton) is quite small, because of the serial correlation in the MCMC iterations.

Parallel to the presentation in the previous section (Table 5.4), Table 6.6 provides log marginal likelihood approximations for each of the fifteen models considered, applied in each case to the artificial data set generated under each of the distributional assumptions. The purpose is to gain some evidence on how well Bayes factors can discriminate among models. When outcomes are fully observed (panels A and B of Table 6.6), the Bayes factors draw very sharp and correct distinctions among models. The results parallel those for the univariate models. The greater contrast arises from the increase in information with the number of dimensions. The lowest odds ratio in favor of a true model in panels A and B is that in favor of the normal multivariate

Table 6.6
Log marginal likelihoods in some multivariate latent linear models with artificial data

Model	Data disturbances		
	Normal	Student-t	Mixed normal
A. Multivariate linear model			
Normal	–4323.7	–5266.4	–4328.1
Student-t	–4335.1	–5031.3	–4162.4
Mixed normal	–4328.1	–5069.5	–2998.2
B. Degenerate multivariate linear model			
Normal	–2205.5	–3009.3	–1672.5
Student-t	–2213.4	–2770.7	–1452.1
Mixed normal	–2211.7	–2803.1	–1298.4
C. Multiple discrete choice model			
Normal	–890.0	–1124.3	–788.8
Student-t	–891.5	–1110.8	–782.6
Mixed normal	–897.2	–1122.0	–780.3
D. Continuous outcome selection model			
Normal	–764.1	–926.4	–1018.1
Student-t	–765.4	–885.1	–989.3
Mixed normal	–768.2	–891.5	–964.1
E. Single discrete choice outcome selection model			
Normal	–1445.4	–1622.2	–730.0
Student-t	–1447.1	–1615.0	–731.2
Mixed normal	–1451.8	–1624.2	–730.0

linear model against the mixed normal linear model, about 80:1. The odds against alternative models that do not nest true models are overwhelming.

For the multiple discrete choice model, no component of the outcome vector $\tilde{\mathbf{y}}_t$ is ever observed directly and the contrasts are much weaker, but the true model is always favored. However, the Student-t distribution provides a close competitor when the true model is normal (4.5:1 in favor of normal) or mixed normal (10:1 in favor of mixed normal). Misspecified models – ones that do not nest the true model – are clearly rejected: for example Student-t is preferred by $7.3 \times 10^6 : 1$ over normal when disturbances are Student-t, and mixed normal by 4900:1 over normal when disturbances are mixed normal.

In the continuous outcome selection model one component of $\tilde{\mathbf{y}}_t$ is observed in somewhat less than half the sample. Again, odds ratios always favor the true model.

For alternative models that nest, or approximately nest, true models odds ratios in favor of the true model are similar here to those seen in all panels. This is not surprising, since the outcome springs from the penalty for more complex models implicit in the prior distribution, and prior distributions are similar across all models. For models that inappropriately restrict the true model, odds ratios for the continuous outcome selection model are substantially higher than those in panels C or E.

In the single discrete choice outcome model, odds ratios favor the correct model in the case of the normal and Student-*t* distributions, but in the case of the mixed normal distribution the odds do not discriminate against the normal distribution. Overall, contrasts are lower for these models than was the case for the multiple discrete choice model. This is unsurprising. In both models the shock distribution is two-dimensional, but whereas in the multiple discrete choice model the data always provide two linear inequality restrictions, in the discrete choice selection model the data provide two such restrictions in fewer than half the observations and only one in the remaining observations.

6.2.7. Prospects for applications and future development

These results indicate that the Bayesian approach to multivariate latent linear models can successfully recover the underlying structure in a variety of situations, when the distribution of the disturbances is substantially more complex than multivariate normal. It can also discriminate between competing, non-nested disturbance distribution specifications.

When the outcome is fully observed, as in the case of the first two of the five examples considered here, maximum likelihood is straightforward. In large sample sizes like those used here the asymptotic sampling distribution provides a good approximation to the posterior distribution, so long as the prior is uninformative relative to the data. When there is actual latency in the outcomes, our Bayesian approach to inference in discrete models appears to be competitive with non-Bayesian alternatives that have been developed. In particular, in the multinomial probit model and the continuous outcome selection model with normal disturbances computation time and complexity appear comparable [Geweke, Keane and Runkle (1994)]. Of course, our approach also permits non-Guassian distributions in these latent variable models, and for these specifications there are presently no widely applied non-Bayesian approaches to inference.

Extension of classical methods to likelihood-based inference for the non-Gaussian distributions taken up here would be awkward. In addition, the applicability of asymptotic theory in typical sample sizes is in doubt, and in the case of the normal mixture distribution the likelihood function is unbounded so long as at least one constituent of $\tilde{\boldsymbol{y}}_t$ is fully observed in part of the sample. The nonparametric approaches to dichotomous discrete choice models discussed in Section 5 – in particular, single-index models – have not been applied to multiple discrete choice problems. We conclude that the approach taken here appears to be on the forefront in the development of flexible models when outcomes are partially observed. Further development and investigation is warranted.

7. Bayesian inference for a dynamic discrete choice model

In this section we present an example of Bayesian inference for a dynamic discrete choice model. The model we consider is dynamic in the sense that current period decisions affect the next period's state variables, and hence the distribution of next period's payoffs. But the inferential procedure we describe here does not require solution of agents' dynamic optimization problem. Instead, we adopt an approach (discussed in Section 3) in which we assume the future components of the value functions lie along a flexible polynomial in the state variables – a polynomial whose arguments are determined by the structure of the model and the laws of motion of the state variables. Then, we form the joint posterior of the polynomial coefficients and the parameters of agents' payoff functions using a Gibbs sampling algorithm. Our discussion is based on Geweke, Houser and Keane (1998), and is based on a model that is very similar to the one analyzed by Keane and Wolpin (1997).

In the model we consider, agents chose among four mutually exclusive alternatives in each of $t = 1, \dots, 40$ periods. The first two alternatives are to work in one of two alternative occupations, the third is to attend school and the fourth to remain home.

One component of the current period payoff in each of the two occupational alternatives is the associated wage: (w_{ijt}), $j = 1, 2$. The log-wage equation is:

$$\begin{aligned}\ln w_{ijt} &= \beta_{oj} + \beta_{1j}X_{i1t} + \beta_{2j}X_{i2t} + \beta_{3j}S_{it} + \beta_{4j}X_{ijt}^2 + \varepsilon_{ijt}\\ &= Y_{ijt}'\beta_j + \varepsilon_{ijt} \quad j = 1, 2,\end{aligned}$$

where Y_{ijt} is the obvious vector, X_{ijt} is periods of experience in occupation j, S_{it} is periods of school completed, and the ε_{ijt} are serially independent productivity shocks, with $(\varepsilon_{i1t}, \varepsilon_{i2t})' \sim N(0, \Sigma_\varepsilon)$. Each occupational alternative also has a stochastic nonpecuniary payoff, v_{ijt}, so the complete current period payoffs are $u_{ijt} = w_{ijt} + v_{ijt}$ $(j = 1, 2)$.

The schooling payoffs include tuition costs. Agents begin with a 10^{th} grade education, and may complete two additional grades without cost. We assume there is a fixed undergraduate tuition rate for attending grades 13 through 16 (α_1), and a fixed graduate tuition rate for each year of schooling beyond 16 (α_2). We assume a "return to school" cost that agents face if they did not choose school the previous period (α_3). Finally, school has both a nonstochastic nonpecuniary benefit (α_0), and a mean zero stochastic nonpecuniary payoff v_{i3t}. Thus we have

$$\begin{aligned}u_{i3t} &= \alpha_0 + \alpha_1\chi(12 \leqslant S_{it} \leqslant 15) + \alpha_2\chi(S_{it} \geqslant 16) + \alpha_3\chi(d_{i,t-1} \neq 3) + v_{i3t}\\ &= \Delta_{it}'\alpha + v_{i3t},\end{aligned}$$

where Δ_{it} is a vector of zeros and ones, according to the values of the indicator functions χ. Lastly, we assume that option 4, "home", has both a nonstochastic nonpecuniary payoff (ϕ), and a mean zero stochastic nonpecuniary payoff v_{i4t}, so that:

$$u_{i4t} = \phi + v_{i4t}.$$

It will be convenient to write $u_{ijt} = \bar{u}_{ijt} + v_{ijt}$ ($j = 1, 4$). We assume the v_{ijt} are serially independent.

The state of the agent at the time of each decision is

$$I_{it} = \left\{ \left(X_{ijt}\right)_{j=1,2}, S_{it}, t, \chi\left(d_{i,t-1} = 3\right), \left(\varepsilon_{ijt}\right)_{j=1,2}, \left(v_{ijt}\right)_{j=1,\ldots,4} \right\},$$

and, in the notation of Section 3, we have:

$$I_{it}^{*} = \left\{ \left(X_{ijt}\right)_{j=1,2}, S_{it}, t, \chi\left(d_{i,t-1} = 3\right) \right\}.$$

The number of "home" choices is excluded from the state-space as it is linearly dependent on the level of education, the period, and experience in the two occupational alternatives.

The value of an alternative is the sum of its current period payoff, and the future component:

$$\begin{aligned} V_{ijt}\left(I_{it}\right) = \bar{u}_{ijt}\left(I_{it}\right) + v_{ijt} + F\left(X_{i1t} + \iota_{1j}, X_{i2t} + \iota_{2j}, S_{it} + \iota_{3j}, t+1, \iota_{3j}\right), \\ j = 1, \ldots, 4, \quad t = 1, \ldots, 40, \end{aligned}$$

where $\iota_{kj} = 1$ if $k = j$ and is zero otherwise. The function F represents agents' forecasts about the effects of their current state and choice on their future payoff stream. The function is fixed across alternatives, so that the forecasts vary across alternatives only because different choices lead to different future states.

The future component's arguments reflect restrictions implied by the model. For instance, because the productivity and preference shocks are serially independent, they contain no information useful for forecasting future payoffs and do not appear in the future component's arguments. Also, many forms of path dependence that are not consistent with the model are ruled out. In particular, given a state, the order in which occupations one and two were chosen does not bear on future payoffs. Accordingly, only their aggregate occurrence enters the future component. It is worthwhile to point out that these restrictions can be viewed as providing logical consistency to the agents' behavior in relation to the model's assumptions.

Since choices depend only on relative alternative values, rather than their levels, we define for $j \in \{1, 2, 3\}$:

$$\begin{aligned} Z_{ijt} &\equiv V_{ijt} - V_{i4t} \\ &= \bar{u}_{ijt} + v_{ijt} + F\left(I_{it}^{*}, j\right) - \bar{u}_{i4t} - v_{i4t} - F\left(I_{it}^{*}, 4\right) \\ &= \tilde{u}_{ijt} + f\left(I_{it}^{*}, j\right) + \eta_{ijt}, \end{aligned}$$

where $\tilde{u}_{ijt} \equiv \bar{u}_{ijt} - \bar{u}_{i4t}, f\left(I_{it}^{*}, j\right) \equiv F\left(I_{it}^{*}, j\right) - F\left(I_{it}^{*}, 4\right)$, and $\eta_{ijt} \equiv v_{ijt} - v_{i4t}$. We assume $\eta_{it} \equiv (\eta_{i1t}, \ldots, \eta_{i3t})' \sim N(0, \Sigma_{\eta})$. Importantly, after differencing, the parameter ϕ of the home payoff is subsumed in f, the relative future component. Clearly, if an alternative's relative future component has an intercept (as each of ours does) then it and the current period return to home cannot be separately identified.

The Z_{it} are unobserved by the econometrician. The econometrician only observes the agents' choices $\{d_{it}\}_{t=1}^{40}$ and, in periods when the agent works, the wage for the chosen alternative. Thus, payoffs are never completely observed, both because wages are censored and because the nonpecuniary components of the payoffs (v_{ijt}) are never observed. Nevertheless, given observed choices and partially observed wages, along with the functional form assumptions about the payoff functions, it is possible to learn both about the future component $F(\cdot)$ and the structural parameters of the payoff functions without making strong assumptions about how agents form expectations. Rather, we simply assume that the future component lies along a polynomial in the relevant state variables. In the Monte Carlo results reported below we used a polynomial of order four. After differencing to obtain $\left\{f\left(I_{it}^{*},j\right)\right\}_{j=1,2,3}$, the polynomial we used contained 53 terms of order three and lower [see Geweke, Houser and Keane (1998, Appendix A) for details].

The relative future component can be expressed:

$$f\left(I_{it}^{*},j\right)=\Psi_{ijt}'\pi \quad j=1,2,3,$$

where Ψ_{ijt} is a vector of functions of state-variables that appear in the equation for $f(j)$ and π is a vector of coefficients common to each choice. Cross-equation restrictions of this type are a consequence of using the same future component function F for each alternative, and reflect the consistency restrictions mentioned earlier.

The first step in a Bayesian analysis of this model via a Gibbs sampler with data augmentation is to form the "complete data" likelihood function. That is, we consider the likelihood function that could be formed if we had data on N individuals observed over 40 periods each, and we observed the value function differences $\boldsymbol{Z}=\left\{\left(Z_{ijt}\right)_{j=1,2,3;\ i=1,N;\ t=1,40}\right\}$ and the complete set of wages $\boldsymbol{W}=\left\{\left(w_{ijt}\right)_{j=1,2;\ i=1,N;\ t=1,40}\right\}$ for all alternatives. This is:

$$\begin{aligned}
&L\left(\boldsymbol{W},\boldsymbol{Z}\mid Y,\Delta,\beta_1,\beta_2,\alpha,\pi,\Sigma_\varepsilon,\Sigma_\eta\right)\\
&\quad\propto\prod_{i,t}\left|\Sigma_\varepsilon\right|^{-1/2}\left(w_{i1t}w_{i2t}\right)^{-1}\exp\left\{-\tfrac{1}{2}\begin{pmatrix}\ln w_{i1t}-Y_{i1t}'\beta_1\\ \ln w_{i2t}-Y_{i2t}'\beta_2\end{pmatrix}'\Sigma_\varepsilon^{-1}\begin{pmatrix}\ln w_{i1t}-Y_{i1t}'\beta_1\\ \ln w_{i2t}-Y_{i2t}'\beta_2\end{pmatrix}\right\}\\
&\qquad\cdot\left|\Sigma_\eta\right|^{-1/2}\exp\left\{-\tfrac{1}{2}\begin{pmatrix}Z_{i1t}-w_{i1t}-\Psi_{i1t}'\pi\\ Z_{i2t}-w_{i2t}-\Psi_{i2t}'\pi\\ Z_{i3t}-\Delta_{it}'\alpha-\Psi_{i3t}'\pi\end{pmatrix}'\Sigma_\eta^{-1}\begin{pmatrix}Z_{i1t}-w_{i1t}-\Psi_{i1t}'\pi\\ Z_{i2t}-w_{i2t}-\Psi_{i2t}'\pi\\ Z_{i3t}-\Delta_{it}'\alpha-\Psi_{i3t}'\pi\end{pmatrix}\right\}\\
&\quad\cdot\chi\big(Z_{ijt}>\max\left\{0,Z_{ikt}\,(k\neq j)\right\}\ \text{if}\ d_{it}=j\\
&\qquad\text{and}\ j\in\{1,2,3\},\ \{Z_{ijt}\}_{j=1,2,3}<0\ \text{otherwise}\big).
\end{aligned}\tag{7.1}$$

We will assume flat priors on all parameters except the two covariance matrices, for which we impose the standard noninformative priors [see Zellner (1971, Section 8.1)]:

$$p\left(\Sigma_\varepsilon^{-1}\right) \propto \left|\Sigma_\varepsilon^{-1}\right|^{-3/2}, \quad p\left(\Sigma_\eta^{-1}\right) \propto \left|\Sigma_\eta^{-1}\right|^{-2}. \tag{7.2}$$

The joint posterior density of the parameters is proportional to the product of Equation (7.1) and the two densities in Equation (7.2). Geweke, Houser and Keane (1998) contains a proof that this joint posterior is finitely integrable.

It is not feasible to construct the parameters' posterior density analytically, because of the high dimensional integration over the unobserved wages and value function differences that is involved. Fortunately, it is possible to simulate draws from the posterior using a Gibbs sampler with data augmentation. As discussed in Section 2, implementing this algorithm requires us to factor the joint posterior into a set of conditional posterior densities, in such a way that each can be drawn from easily. Then, we cycle through these conditionals, drawing a block of parameters from each in turn. As the number of cycles grows large, the parameter draws so obtained converge in distribution to their respective marginal posteriors. Our Gibbs sampling-data augmentation algorithm consists of six steps or "blocks", which are as follows:

Step 1. Draw the value function differences $\{Z_{ijt}, i = 1, N; j = 1, 2, 3; t = 1, 40\}$.
Step 2. Draw the unobserved wages $\{w_{ijt}$ when $d_{it} \neq j,\ (j = 1, 2)\}$.
Step 3. Draw the log-wage equation coefficients β_j.
Step 4. Draw the log-wage equation error-covariance matrix Σ_ε.
Step 5. Draw the future component parameters π and school payoff parameters α.
Step 6. Draw the nonpecuniary payoff covariance matrix Σ_η.

We next briefly describe how each step was carried out. Additional detail can be found in Geweke, Houser and Keane (1998).

Step 1. Taking everything else in the model as given, it is evident from Equation (7.1) that the conditional distribution of a single Z_{ijt} is truncated Gaussian. There are three ways in which the distribution might be truncated. In case 1: Z_{ijt} is a value function difference for the chosen alternative. In this case we draw

$$Z_{ijt} > \max\left\{0, (Z_{ikt})_{\substack{k \in \{1,2,3\} \\ k \neq j}}\right\}.$$

In case 2: Z_{ijt} is not associated with the chosen alternative, and "home" was not chosen. In this case, we draw $Z_{ijt} < Z_{id_{it}t}$. In case 3: "home" was chosen. In this case, we draw $Z_{ijt} < 0$. We draw from the univariate, truncated Gaussian distributions using standard inverse CDF methods.

Step 2. Drawing unobserved wages is the most time consuming part of the algorithm. Suppose w_{i1t} is unobserved. Its density, conditional on every other wage, future component difference and parameter being known, is from Equation (7.1) given by:

$$g\left(\tilde{w}_{i1t}|\cdot\right) \propto \tilde{w}_{i1t}^{-1} \exp\left\{-\tfrac{1}{2}\begin{pmatrix} \ln \tilde{w}_{i1t} - Y_{i1t}'\beta_1 \\ \ln w_{i2t} - Y_{i2t}'\beta_2 \end{pmatrix}' \Sigma_\varepsilon^{-1} \begin{pmatrix} \ln \tilde{w}_{i1t} - Y_{i1t}'\beta_1 \\ \ln w_{i2t} - Y_{i2t}'\beta_2 \end{pmatrix}\right\}$$
$$\exp\left\{-\tfrac{1}{2}\begin{pmatrix} Z_{i1t} - \tilde{w}_{i1t} - \Psi_{i1t}'\pi \\ Z_{i2t} - w_{i2t} - \Psi_{i2t}'\pi \\ Z_{i3t} - \Delta_{it}'\alpha - \Psi_{i3t}'\pi \end{pmatrix}' \Sigma_\eta^{-1} \begin{pmatrix} Z_{i1t} - \tilde{w}_{i1t} - \Psi_{i1t}'\pi \\ Z_{i2t} - w_{i2t} - \Psi_{i2t}'\pi \\ Z_{i3t} - \Delta_{it}'\alpha - \Psi_{i3t}'\pi \end{pmatrix}\right\}. \tag{7.3}$$

This distribution is nonstandard as wages enter in both logs and levels. Nevertheless, it is straightforward to sample from this distribution using rejection methods [see Geweke (1996) for a discussion of rejection sampling]. In brief, we first draw a candidate wage w^c from the distribution implied by the first exponential term of Equation (7.3), so that $\ln w^c \sim N\left(Y_{i1t}'\beta_1 + \lambda_{it}, \sigma_*^2\right)$, where $\lambda_{it} \equiv \Sigma_\varepsilon(1,2)\varepsilon_{i2t}/\Sigma_\varepsilon(2,2)$ and $\sigma_*^2 \equiv \Sigma_\varepsilon(1,1)\left(1 - \left(\Sigma_\varepsilon(1,2)^2/\Sigma_\varepsilon(1,1)\,\Sigma_\varepsilon(2,2)\right)\right)$. This draw is easily accomplished, and w^c is found by exponentiating. We accept the draw with probability equal to the second exponential term in Equation (7.3), when evaluated at $\tilde{w}_{i1t} = w^c$, divided by the conditional maximum of this term over $\tilde{w}_{i1t}$. If the draw is accepted then the unobserved w_{i1t} is set to w^c. Otherwise, the process is repeated until a draw is accepted.

Step 3. Given all wages, value function differences, and other parameters, the density of (β_1, β_2) is:

$$g\left(\beta_1, \beta_2\right) \propto \exp\left\{-\tfrac{1}{2}\begin{pmatrix} \ln w_{i1t} - Y_{i1t}'\beta_1 \\ \ln w_{i2t} - Y_{i2t}'\beta_2 \end{pmatrix}' \Sigma_\varepsilon^{-1} \begin{pmatrix} \ln w_{i1t} - Y_{i1t}'\beta_1 \\ \ln w_{i2t} - Y_{i2t}'\beta_2 \end{pmatrix}\right\},$$

so that (β_1, β_2) is distributed according to a multivariate normal. In particular, it is easy to show that

$$\beta \sim N\left[\left(\boldsymbol{Y}'\Sigma^{-1}\boldsymbol{Y}\right)^{-1}\boldsymbol{Y}'\Sigma^{-1}\ln\boldsymbol{W}, \left(\boldsymbol{Y}'\Sigma^{-1}\boldsymbol{Y}\right)^{-1}\right],$$

where $\beta \equiv (\beta_1', \beta_2')'$, $\Sigma = \Sigma_\varepsilon \otimes I_{\mathrm{NT}}$, $\boldsymbol{Y} = \begin{bmatrix} \boldsymbol{Y}_1 & 0 \\ 0 & \boldsymbol{Y}_2 \end{bmatrix}$ and $\ln\boldsymbol{W} = [\ln\boldsymbol{W}'_1, \ln\boldsymbol{W}'_2]'$, where $\boldsymbol{Y}_1$ is the regressor matrix for the first log-wage equation naturally ordered through all individuals and periods, and similarly for $\boldsymbol{Y}_2$, $\boldsymbol{W}_1$, and $\boldsymbol{W}_2$. It is straightforward to draw β from this multivariate normal density.

Step 4. With everything else known, Σ_ε^{-1} has a Wishart distribution. Specifically,

$$\Sigma_\varepsilon^{-1} \sim W\left(SST_\varepsilon, N \cdot T\right),$$

where $SST_\varepsilon = \sum_{i,t}(\varepsilon_{i1t}\varepsilon_{i2t})'(\varepsilon_{i1t}\varepsilon_{i2t})$, and $\varepsilon_{ijt} = \ln w_{ijt} - Y_{ijt}'\beta_j$. It is easy to draw from the Wishart and then invert the 2 × 2 matrix to obtain the new Σ_ε.

Step 5. It is convenient to draw both the future component π parameters and the parameters α of the school payoff jointly. Since the future component for school contains an intercept, it and the constant in α cannot be separately identified. Hence, we omit α_0 as well as the first element from each Δ_{it}. Define the vector $\pi^* \equiv [\pi', \alpha']'$, where $\alpha = (\alpha_1, \alpha_2, \alpha_3)'$ and define $\Psi^*_{ijt} \equiv \left[\Psi'_{ijt}, 0'_3\right]'$ $(j = 1, 2)$, and $\Psi'_{i3t} = \left[\Psi'_{i3t}, \Delta'_{it}\right]'$. Note that π^* and the Ψ^*_{ijt} are 56-vectors. Then define $\Psi_k = \left[\Psi^*_{1k1}, \Psi^*_{1k2}, \ldots, \Psi^*_{Nk,T-1}, \Psi^*_{NkT}\right]$ and set $\Psi = [\Psi_1, \Psi_2, \Psi_3]'$ so that Ψ is a $(3 \cdot NT \times 56)$ stacked-regressor matrix. Similarly, define the corresponding $3 \cdot NT$-vector Γ by:

$$\Gamma = \left(\{Z_{i1t} - w_{i1t}\}'_{i=1,N;\ t=1,40}\ \{Z_{i2t} - w_{i2t}\}'_{i=1,N;\ t=1,40}\ \{Z_{i3t}\}'_{i=1,N;\ t=1,40}\right)'.$$

It is immediate from Equation (7.1), in which π^* enters only through the second exponential expression, that, conditional on everything else in the model known, π^* has a multivariate normal density given by:

$$\pi^* \sim N\left[\left(\Psi'\Omega^{-1}\Psi\right)^{-1}\Psi'\Omega^{-1}\Gamma, \left(\Psi'\Omega^{-1}\Psi\right)^{-1}\right],$$

where $\Omega = \Sigma_\eta \otimes I_{NT}$. It is straightforward to draw from this distribution using a standard, multivariate normal random number generator.

Step 6. With everything else known the distribution of Σ_η^{-1} is Wishart. Specifically,

$$\Sigma_\eta^{-1} \sim W\left(SST_\eta, N \cdot T\right),$$

where $SST_\eta = \sum_{i,t}(\eta_{i1t}\eta_{i2t}\eta_{i3t})'(\eta_{i1t}\eta_{i2t}\eta_{i3t})$. It is easy to draw from this distribution and then invert the 3×3 matrix to obtain the new Σ_η.

To investigate the performance of this algorithm, we generated five artificial data sets with $N = 2000$ each, using the true parameter values listed in column 2 of Table 7.1. We generated data by solving the dynamic optimization problem "exactly" under rational expectations given these parameter values, and forming the optimal decision rule. Thus, the point of this experiment is to gauge if our method will generate reliable inferences about the structural parameters of the payoff functions even when the polynomial approximation to the future component is misspecified.

Starting from an initial guess of the model parameters[34], we ran the Gibbs algorithm for 40 000 cycles. Visual inspection of graphs of the draw sequences, as

[34] We chose to set the initial log-wage equation β equal to the value from an OLS regression on observed wages. The diagonal elements of Σ_ε were set to the variance of observed log-wages, while the off-diagonal elements were set to zero. The school payoff parameters were all initialized at zero. All of the future component's π values were also started at zero, except for the alternative-specific intercepts. The intercepts for alternatives one, two and three were initialized with −5000, −10 000 and 20 000, respectively (which were not the true values). These values were chosen with an eye toward matching aggregate choice frequencies in each alternative. Finally, we also chose an arbitrary initialization for the Σ_η covariance matrix. We set all off-diagonal elements to zero, and set each diagonal element to 5×10^8. We used large starting variances because doing so increases the size of the initial Gibbs steps, and seems to improve the rate of convergence of the algorithm.

well as application of the split sequence diagnostic suggested by Gelman (1996) – which compares variability of the draws across subsequences – suggest that the algorithm converged for all five artificial data sets. We then used the last 15 000 draws from each run to simulate the posterior. Table 7.1 reports the posterior means and posterior standard deviations of the simulated posterior distribution for each structural parameter of the current payoff functions (the future component π parameters are not reported).

The performance of the algorithm is quite impressive. In almost all cases, the posterior means of the wage function parameters deviate only slightly from the true values in percentage terms. Also, the posterior standard deviations are in most cases quite small, suggesting that the data contain a great deal of information about these structural parameters – even without imposing the assumption that agents form the future component "optimally". Finally, despite that fact that the posterior standard deviations are quite small, the posterior means are rarely more than two posterior deviations away from the true values. The school payoff parameters are not pinned down so well as the wage equation parameters, which is not surprising given that school payoffs are never observed.

These findings are related to those of Lancaster (1997), who considered Bayesian inference in the stationary job search model. He found that if the reservation wage is treated as a free parameter, rather than imposing that it is set "optimally" (as dictated by the offer wage function, offer arrival rate, unemployment benefit and discount rate), there is little loss of information about the structural parameters of the offer wage functions. (As in our example, however, identification of the discount factor is lost.) The stationary job search model considered by Lancaster (1997) has the feature that the future component is a constant (i.e., it is not a function of state variables). Our procedure of treating the future component as a polynomial in state variables can be viewed as extending Lancaster's approach to a much more general class of models.

The results of Table 7.1 indicate that in a case where agents form the future component optimally, we can still obtain reliable and precise inferences about structural parameters of the current payoff functions using a simplified and misspecified model that says the future component is a simple 4^{th} order polynomial in the state variables. But we are also interested in how well our method approximates the decision rule used by the agents. In Table 7.2 we consider an experiment in which we use the posterior means for the parameters π that characterize how agents form expectations to form an estimate of agents' decision rule. We then simulate 5 new artificial data sets, using the exact same draws for the current period payoffs as were used to generate the original 5 artificial data sets. The only difference is that the estimated future component is substituted for the true future component in forming the decision rule. The results in Table 7.2 indicate that the mean wealth losses from using the estimated decision rule range from five-hundredths to three-tenths of one percent. The percentage of choices that agree between agents who use the optimal versus the approximate rules ranges from 89.8% to 93.5%. These results suggest that

Table 7.1
Descriptive statistics for final 15 000 Gibbs sampler parameter draws for several different data sets generated using true future component

Parameter	True	Data Set 1		Data Set 2		Data Set 3		Data Set 4		Data Set 5	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Occ. 1 intercept	9.00000	9.01342	0.00602	9.00471	0.00527	9.01436	0.00584	9.01028	0.00593	9.00929	0.00550
Occ. 1 own experience	0.05500	0.05427	0.00073	0.05489	0.00071	0.05384	0.00072	0.05394	0.00072	0.05410	0.00071
Occ. 2 experience	0.00000	0.00111	0.00093	0.00092	0.00114	0.00078	0.00126	0.00107	0.00100	0.00051	0.00093
Education	0.05000	0.04881	0.00118	0.05173	0.00126	0.04869	0.00129	0.04961	0.00123	0.05067	0.00124
Occ. 1 exp. squared	–0.00025	–0.00023	0.00002	–0.00025	0.00002	–0.00023	0.00002	–0.00022	0.00002	–0.00023	0.00002
Occ. 1 error SD	0.40000	0.39740	0.00200	0.39870	0.00200	0.39850	0.00200	0.39730	0.00200	0.39740	0.00200
Occ. 2 intercept	8.95000	8.90720	0.01704	8.98989	0.01970	8.93943	0.01850	8.93174	0.01649	8.94097	0.01410
Occ. 2 own experience	0.04000	0.04093	0.00037	0.03967	0.00037	0.03955	0.00038	0.04001	0.00037	0.04060	0.00039
Occ. 2 experience	0.06000	0.06087	0.00178	0.05716	0.00190	0.06200	0.00201	0.06211	0.00179	0.05880	0.00157
Education	0.07500	0.07822	0.00166	0.07338	0.00171	0.07579	0.00165	0.07743	0.00167	0.07613	0.00159
Occ. 2 exp. squared	–0.00090	–0.00087	0.00008	–0.00081	0.00008	–0.00098	0.00008	–0.00101	0.00008	–0.00084	0.00007
Occ. 2 error SD	0.40000	0.40850	0.00300	0.39680	0.00300	0.40390	0.00300	0.40240	0.00300	0.39720	0.00300
Error correlation	0.50000	0.51690	0.02300	0.60680	0.02900	0.48420	0.04400	0.52110	0.03500	0.48750	0.02800
Undergraduate tuition	–5000	–2261	313	–2937	358	–3407	371	–3851	426	–3286	448
Graduate tuition	–15000	–10092	1046	–10788	141	–11983	1188	–10119	1380	–11958	1823
Return cost	–15000	–14032	482	–16014	431	–16577	500	16168	662	–18863	1065

continued on next page

Table 7.1, *continued*

Parameter	True	Data Set 1		Data Set 2		Data Set 3		Data Set 4		Data Set 5	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Preference shock SD											
Occ. 1	9082.95	10634.90	423.85	10177.24	165.11	11438.63	438.72	9973.32	371.64	9071.29	509.80
Occ. 2	9082.95	9436.10	372.86	12741.02	405.25	11432.19	287.69	9310.37	718.15	7770.66	555.39
Occ. 3	11821.59	11450.65	338.28	12470.12	259.81	13999.95	351.33	13183.33	471.47	13897.62	533.67
Preference shock corr.											
Occ. 1 with occ. 2	0.89	0.93	0.01	0.98	0.00	0.94	0.01	0.91	0.02	0.86	0.03
Occ. 1 with occ. 3	0.88	0.89	0.01	0.88	0.01	0.90	0.01	0.88	0.01	0.88	0.01
Occ. 2 with occ. 3	0.88	0.87	0.01	0.90	0.01	0.90	0.01	0.89	0.02	0.89	0.02

Table 7.2
Wealth loss when posterior polynomial approximation is used in place of true future component[1]

Data set	True EMAX[2]	Posterior EMAX						
	Mean PV of payoffs[3]	Mean PV of payoffs	Mean dollar equivalent loss	Mean loss (%)	Aggregate choice agreement	0–35 agreements (%)	36–39 agreements (%)	Choosing same path (%)
1	356796.40	356133.68	662.71	0.19%	90.80%	34.25%	42.65%	23.10%
2	356326.99	355836.27	490.72	0.14%	91.34%	33.00%	44.00%	23.00%
3	355796.91	354746.00	1050.91	0.30%	89.79%	39.00%	38.95%	22.05%
4	355802.85	355449.73	353.12	0.10%	93.48%	24.60%	38.45%	36.95%
5	355660.94	355484.65	176.29	0.05%	93.18%	24.95%	30.50%	44.55%

[1] Polynomial parameter values are set to the mean of their respective empirical posterior distributions from data set 5.

[2] Each simulation includes 2000 agents that live for exactly 40 periods.

[3] "Mean PV of payoffs" is the equal-weight sample average of discounted streams of ex-post lifetime payoffs.

our estimated polynomial approximations to the optimal decision rule are indeed reasonably accurate.

Appendix A. The full univariate latent linear model

Core model. The core univariate latent linear model (ULLM) is

$$\tilde{y}_t = \alpha' \tilde{\boldsymbol{z}}_t + \beta' \boldsymbol{x}_t + \varepsilon_t.$$

The subscript t indexes observations. The $k \times 1$ vector $\boldsymbol{x}_t$ consists of observed covariates, $k > 0$. The $m \times 1$ vector $\tilde{\boldsymbol{z}}_t$ consists of latent variables, $m \geqslant 0$. The $m \times 1$ parameter vector α and the $k \times 1$ parameter vector β are unknown. The outcome variable $\tilde{y}_t$ may be either observed, or latent subject to known restrictions. In the core ULLM the independently and identically distributed (i.i.d.) disturbances ε_t have a scale mixture of normals distribution, with representation

$$\varepsilon_t = \sigma_{(t)} \eta_t. \tag{A.1}$$

In Equation (A.1), $\sigma_{(t)}$ is a strictly positive, i.i.d. random variable conditional on $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T$, and $(\eta_1, \ldots, \eta_T | \sigma_{(1)}, \ldots, \sigma_{(T)}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_T) \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_T)$. Hence

$$p\left(\{\tilde{y}_t\}_{t=1}^T \mid \left\{\boldsymbol{x}_t, \tilde{\boldsymbol{z}}_t, \sigma_{(t)}^2\right\}_{t=1}^T\right) = (2\pi\sigma^2)^{-T/2} \left[\prod_{t=1}^T \left(\sigma_{(t)}^2\right)^{-1/2}\right] \cdot \exp\left\{-\sum_{t=1}^T \left[\left(\tilde{y}_t - \alpha'\tilde{\boldsymbol{z}}_t - \beta'\boldsymbol{x}_t\right)^2 / 2\sigma^2\sigma_{(t)}^2\right]\right\}. \tag{A.2}$$

Let $\tilde{\boldsymbol{y}}' = (\tilde{y}_1, \ldots, \tilde{y}_T), \boldsymbol{X}' = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T], \tilde{\boldsymbol{Z}}' = [\tilde{\boldsymbol{z}}_1, \ldots, \tilde{\boldsymbol{z}}_T]$, and let $\boldsymbol{Q}$ be a $T \times T$ diagonal matrix with entry $\sigma_{(t)}^2$ in position (t, t). Then an alternate representation of Equation (A.2) is

$$p\left(\tilde{\boldsymbol{y}} | \boldsymbol{X}, \tilde{\boldsymbol{Z}}, \boldsymbol{Q}\right) = (2\pi\sigma^2)^{-T/2} \left|\boldsymbol{Q}\right|^{-1/2} \exp\left[-\left(\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{Z}}\alpha - \boldsymbol{X}\beta\right)' \boldsymbol{Q} \left(\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{Z}}\alpha - \boldsymbol{X}\beta\right) / 2\sigma^2\right]. \tag{A.3}$$

Distributional assumptions. The core ULLM yields a rich variety of distributions as specific cases. The treatment here considers three.

When $m = 0$ and $\sigma_{(t)}^2 \equiv 1$, then $\tilde{\boldsymbol{y}} | \boldsymbol{X} \sim N\left(\boldsymbol{X}\beta, \sigma^2 \boldsymbol{I}_T\right)$. If $\tilde{\boldsymbol{y}}$ is observed this is the textbook normal linear regression model. Other assumptions about the observation of $\tilde{\boldsymbol{y}}$, discussed subsequently, yield the conventional binomial probit and normal censored regression models.

When $m = 0$ and $\lambda\sigma_{(t)}^{-2}|\boldsymbol{X} \overset{\text{i.i.d.}}{\sim} \chi^2(\lambda)$, then

$$p\left(\sigma_{(t)}^2|\lambda,\boldsymbol{X}\right) = \left[2^{\lambda/2}\Gamma(\lambda/2)\right]^{-1}\lambda^{\lambda/2}\left(\sigma_{(t)}^2\right)^{-(\lambda+2)/2}\exp\left(-\lambda/2\sigma_{(t)}^2\right). \tag{A.4}$$

In this case,

$$\varepsilon_t|(\lambda,\boldsymbol{X}) \overset{\text{i.i.d.}}{\sim} t(\boldsymbol{0},\sigma^2;\,\lambda).$$

In the normal mixture model the number of components is $m \geqslant 2$. The random vector $\left(\tilde{\boldsymbol{z}}_t', \sigma_{(t)}^2\right)$ is i.i.d., conditional on $\boldsymbol{X}$. To describe its distribution let $s(t) \in \{1, \ldots, m\}$ be a latent, scalar state index, distributed i.i.d. multinomial conditional on $\boldsymbol{X}$, with

$$P[s(t) = j|\boldsymbol{X}] = p_j; \quad \sum_{j=1}^{m} p_j = 1. \tag{A.5}$$

Corresponding to each state j there is a positive parameter σ_j^2, and $\sigma_{(t)}^2 = \sigma_{s(t)}^2$. The state index $s(t)$ also determines $\tilde{\boldsymbol{z}}_t$: $\tilde{z}_{tj} = \delta_{s(t),j}$, where $\delta_{u,v}$ is the Kronecker delta function $\delta_{u,v} = 1$ if $u = v$ and $\delta_{u,v} = 0$ if $u \neq v$. Then

$$p(\tilde{y}_t|\boldsymbol{x}_t, s(t) = j) = \left(2\pi\sigma^2\sigma_j^2\right)^{-1/2}\exp\left[-\left(\tilde{y}_t - \alpha_j - \beta'\boldsymbol{x}_t\right)^2/2\sigma^2\sigma_j^2\right], \tag{A.6}$$

so that if $s(t) = j$, then $\tilde{y}_t = \beta'\boldsymbol{x}_t + u_t$ with $u_t \sim N(\alpha_j, \sigma^2\sigma_j^2)$. Thus the disturbances u_t follow a full discrete normal mixture distribution.

Observed outcomes. Corresponding to the latent outcome $\tilde{y}_t$ is an observed, set-valued outcome y_t such that $\tilde{y}_t \in y_t$. The core assumption about y_t is that $p(y_t|\tilde{y}_t,\boldsymbol{x}_t) = p(y_t|\tilde{y}_t)$: that is, the value of y_t is determined solely by $\tilde{y}_t$. Thus the joint conditional distribution of $\tilde{y}_t$ and y_t in the ULLM is

$$p(\tilde{y}_t|\boldsymbol{x}_t)\,p(y_t|\tilde{y}_t,\boldsymbol{x}_t) = p(\tilde{y}_t|\boldsymbol{x}_t)\,p(y_t|\tilde{y}_t) = p(\tilde{y}_t|\boldsymbol{x}_t)\,\chi_{y_t}(\tilde{y}_t),$$

where the indicator function $\chi_S(z) = 1$ if $z \in S$, $\chi_S(z) = 0$ if $z \notin S$.

In the linear model y_t is the singleton $y_t = \tilde{y}_t$. In the dichotomous choice model $y_t = (-\infty, 0]$ if choice 1 is observed and $y_t = (0, \infty)$ if choice 2 is observed. In the censored regression model suppose $\tilde{y}_t$ is observed if and only if $\tilde{y}_t \geqslant c$. Then $y_t = \tilde{y}_t \cdot \chi_{[c,\infty)}(\tilde{y}_t) + (-\infty, c) \cdot \chi_{(-\infty,c)}(\tilde{y}_t)$.

All of these models are special instances of the cases $y_t = [c_t, d_t]$, $y_t = (c_t, d_t]$ or $y_t = [c_t, d_t)$. Correspondingly, conditional on $\tilde{y}_t$

$$P(y_t = (c_t, d_t]) = P(y_t = [c_t, d_t)) = P(y_t = [c_t, d_t]) = \chi_{(c_t, d_t]}(\tilde{y}_t), \tag{A.7}$$

with the understanding that c_t and d_t are extended real numbers, and the recognition that $P(\tilde{y}_t = y^*|\boldsymbol{x}_t) = 0$ for any single point y^*, including $y^* = +\infty$ and $y^* = -\infty$.

These instances are of course restrictions on the assumption that y_t is set valued, but they include most cases of interest and simplify both software and the representation of observed data.

Prior distributions. Using Bayes factors to compare different models for the same data requires proper prior distributions. The prior distributions should be chosen so that it is easy to make the prior information about the same for the common parts of alternative models.

The coefficient vector β and disturbance scale parameter σ are common to all models. The benchmark prior distribution for β is $\beta \sim N\left(\underline{\beta}, \underline{\boldsymbol{H}}_\beta^{-1}\right)$:

$$p(\beta) = (2\pi)^{-k/2}\left|\underline{\boldsymbol{H}}_\beta\right|^{1/2}\exp\left[-\tfrac{1}{2}\left(\beta-\underline{\beta}\right)'\underline{\boldsymbol{H}}_\beta\left(\beta-\underline{\beta}\right)\right]. \tag{A.8}$$

The hyperparameters are the mean $\underline{\beta} \in \mathfrak{R}^k$ and the positive definite precision matrix $\underline{\boldsymbol{H}}_\beta$. The prior distribution of σ^2 is inverted gamma, $\underline{s}^2/\sigma^2 \sim \chi^2(\underline{\nu})$:

$$p\left(\sigma^2\right) = \left[2^{\underline{\nu}/2}\Gamma(\underline{\nu}/2)\right]^{-1}\left(\underline{s}^2\right)^{\underline{\nu}/2}\left(\sigma^2\right)^{-(\underline{\nu}+2)/2}\exp\left(-\underline{s}^2/2\sigma^2\right). \tag{A.9}$$

The hyperparameters are the scaling factor $\underline{s}^2 \in \mathfrak{R}^+$ and the degrees of freedom parameter $\underline{\nu} \in \mathfrak{R}^+$.

For the Student-t model the benchmark prior distribution for the degrees of freedom parameter λ is $\lambda \sim \exp(\underline{\lambda})$:

$$p(\lambda) = \underline{\lambda}^{-1}\exp(-\lambda/\underline{\lambda}). \tag{A.10}$$

The hyperparameter $\underline{\lambda} \in \mathfrak{R}^+$ is the mean of the exponential distribution.

The normal mixture model for the disturbances has three components. The multinomial distribution of the state index involves the probabilities $p_1, \ldots, p_m$, $\sum_{j=1}^m p_j = 1$. The benchmark prior distribution is Dirichlet (multivariate beta) with hyperparameters $r_1, \ldots, r_m$:

$$p(\boldsymbol{p}) = \left[\Gamma\left(\sum_{j=1}^m r_j\right)\Big/\prod_{j=1}^m \Gamma\left(r_j\right)\right]\prod_{j=1}^m p_j^{r_j-1}. \tag{A.11}$$

Conditional on one of the m states, say j, the distribution is $N(\alpha_j, \sigma^2\sigma_j^2)$. The benchmark prior distribution for the second component of the normal mixture model, the variance scaling parameters σ_j^2, consists of the m inverted gamma components $\underline{s}_j^2/\sigma_j^2 \sim \chi^2(\underline{\nu}_j)$. These are subject to the restrictions $\sigma_1^2 > \cdots > \sigma_m^2$ but otherwise independent. Thus

$$\begin{aligned} p\left(\sigma_1^2, \ldots, \sigma_m^2\right) &= c\left(\underline{s}_1^2, \ldots, \underline{s}_m^2; \underline{\nu}_1, \ldots, \underline{\nu}_m\right) \\ &\quad \cdot \prod_{j=1}^m\left\{\left[2^{\underline{\nu}_j/2}\Gamma\left(\underline{\nu}_j/2\right)\right]^{-1}\left(\underline{s}_j^2\right)^{\underline{\nu}_j/2}\left(\sigma_j^2\right)^{-(\underline{\nu}_j+2)/2}\exp\left(-\underline{s}_j^2/2\sigma_j^2\right)\right\} \\ &\quad \cdot \chi_{\left\{\sigma_1^2>\cdots>\sigma_m^2\right\}}\left(\sigma_1^2, \cdots, \sigma_m^2\right). \end{aligned} \tag{A.12}$$

The constant $c(\underline{s}_1^2, \ldots, \underline{s}_m^2; \underline{\nu}_1, \ldots, \underline{\nu}_m)$ is the inverse of $P(\sigma_1^2 > \cdots > \sigma_m^2)$ for the independent distributions $\underline{s}_j^2/\sigma_j^2 \sim \chi^2(\underline{\nu}_j)$.

The third component of the normal mixture distribution, $\alpha' = (\alpha_1, \ldots, \alpha_m)$, is multivariate normal, $\alpha|\sigma \sim N(\mathbf{0}, \sigma^2 \underline{\boldsymbol{H}}_\alpha^{-1})$:

$$p(\alpha) = (2\pi\sigma^2)^{-m/2} \left|\underline{\boldsymbol{H}}_\alpha\right|^{1/2} \exp\left(-\alpha' \underline{\boldsymbol{H}}_\alpha \alpha / 2\sigma^2\right). \tag{A.13}$$

Existence of the posterior distribution and moments. Given proper priors and a bounded likelihood function the posterior kernel is integrable and the posterior distribution exists. In the ULLM prior are always proper. The likelihood function is bounded, except when the disturbances are mixed normal and at least some of the $\tilde{y}_t$ are not latent (i.e., $c_t = d_t$ for at least some t). The problem is that α_j can be chosen to make $y_{t^*} = \alpha_j + \beta' \boldsymbol{x}_t$ for some t^*, and then $\sigma_j^2 \to 0$ drives the likelihood to $+\infty$. There exist separated continua of points where the likelihood function is unbounded. Thus maximum likelihood estimation for the normal mixture model with at least one fully observed outcome $y_t = \tilde{y}_t$ is precluded[35].

In this troublesome case the product of prior and data density is still integrable. To see that this is so, without loss of generality consider the instance in which all $\tilde{y}_t$ are observed. Conditional on the state assignments $s(1), \ldots, s(T)$, the posterior density kernel for α, β, σ^2, and $\sigma_j^2 (j = 1, \ldots, m)$ is the product of Equations (A.6, A.8, A.9, A.12 and A.13). This product is bounded above by a fixed multiple of

$$\left(\sigma^2\right)^{-(T+\underline{\nu}+2)/2} \exp\left(-\underline{s}^2/2\sigma^2\right) \prod_{j=1}^{m} \left(\sigma_j^2\right)^{-(T+\underline{\nu}_j+2)/2} \exp\left(-\underline{s}_j^2/2\sigma_j^2\right), \tag{A.14}$$

where $T_j = \sum_{t=1}^{T} \delta_{j,s(t)}$, the number of observations assigned to state j. Expression (A.14) is the product of the density kernels of $\underline{s}^2/\sigma^2 \sim \chi^2(T + \underline{\nu})$ and $\underline{s}_j^2/\sigma_j^2 \sim \chi^2$ $(T_j + \underline{\nu}_j)$ $(j = 1, \ldots, m)$, and is therefore finitely integrable for given $T_1, \ldots, T_m$. Since the number of possible combinations of the T_j is finite the full posterior kernel, unconditional on state assignments, is also finitely integrable.

Given the existence of the posterior density, all posterior moments of any function bounded below and above exist. The posterior moment of an unbounded function of interest exists if the corresponding prior moment exists (a condition that is typically easy to check) and the data density is bounded.

To verify the existence of posterior moments in the case of the normal mixture model with observed outcomes $y_t = \tilde{y}_t$, express Equation (A.14) as the product

$$\left(\sigma^2\right)^{-(\underline{\nu}-\varepsilon)/2} \exp\left[-\left(\underline{s}^2 - \tau^2\right)/2\sigma^2\right] \prod_{j=1}^{m} \left(\sigma_j^2\right)^{-(T+\underline{\nu}_j-\varepsilon)/2} \exp\left[-\left(\underline{s}_j^2 - \tau^2\right)/2\sigma_j^2\right] \tag{A.15}$$

[35] For an early discussion of this problem see Kiefer and Wolfowitz (1956).

$$\cdot \left(\sigma^2\right)^{-(T+\varepsilon+2)/2}\exp(-\tau^2/2\sigma^2)\prod_{j=1}^{m}\left(\sigma_j^2\right)^{-(T_j+\varepsilon+2)/2}\exp\left(-\tau_j^2/2\sigma_j^2\right), \qquad \text{(A.16)}$$

for some positive ε and τ^2. Expression (A.16) is the product of density kernels of the distributions $\tau^2/\sigma^2 \sim \chi^2(T+\varepsilon)$ and $\tau_j^2/\sigma_j^2 \sim \chi^2(T_j+\varepsilon)$ $(j = 1, \ldots, m)$ and is therefore finitely integrable. Expression (A.15) is the product of density kernels of the distributions

$$\left(\underline{s}^2-\tau^2\right)/\sigma^2 \sim \chi^2(\underline{\nu}-2-\varepsilon), \left(\underline{s}_j^2-\tau^2\right)/\sigma^2 \sim \chi^2\left(\underline{\nu}_j-2-\varepsilon\right) \quad (j=1,\ldots,m). \qquad \text{(A.17)}$$

Hence a sufficient condition for the existence of a posterior moment in the normal mixture model with observed outcomes $y_t = \tilde{y}_t$ is that the corresponding prior moment exists when the prior distributions of σ^2 and σ_j^2 $(j = 1, \ldots, m)$ are changed to Equation (A.17). For example, $\underline{\nu}_j > 2 + 2q$ is sufficient for the existence of $E\left[\left(\sigma_j^2\right)^q \mid \boldsymbol{y}, \boldsymbol{X}\right]$.

Inference in the ULLM. There are eight groups of parameters or latent variables in the model: (α,β); σ^2; $\sigma_{(t)}^2$ $(t = 1, \ldots, T)$; λ; $s(t)$ $(t = 1, \ldots, T)$ and $\tilde{\boldsymbol{Z}}$; $\boldsymbol{p}$; σ_j^2 $(j = 1, \ldots, m)$; and $\tilde{y}_t$ $(t = 1, \ldots, T)$. These groups organized the Gibbs sampling MCMC algorithm outlined in Section 5.1.

For the group (α,β), let $\gamma' = (\alpha', \beta')$, $\tilde{\boldsymbol{W}} = \left[\tilde{\boldsymbol{Z}} \vdots \boldsymbol{X}\right]$, $\underline{\gamma}' = (\boldsymbol{0}' \, \underline{\beta}')$, and

$$\underline{\boldsymbol{H}}_\gamma = \begin{bmatrix} \sigma^{-2}\underline{\boldsymbol{H}}_\alpha & \boldsymbol{0} \\ \boldsymbol{0} & \underline{\boldsymbol{H}}_\beta \end{bmatrix}.$$

In the normal and Student-t models, $\gamma = \beta$, $\tilde{\boldsymbol{W}} = \boldsymbol{X}$, $\underline{\gamma} = \underline{\beta}$, and $\underline{\boldsymbol{H}}_\gamma = \underline{\boldsymbol{H}}_\beta$. The kernel of the conditional posterior density is the product of Equations (A.2, A.8 and A.13), from which the conditional posterior distribution is $\gamma \sim N(\bar{\gamma}, \bar{\boldsymbol{H}}_\gamma)$ with $\bar{\boldsymbol{H}}_\gamma = \underline{\boldsymbol{H}}_\gamma + \tilde{\boldsymbol{W}}'(\sigma^2\boldsymbol{Q})^{-1}\tilde{\boldsymbol{W}}$ and $\bar{\gamma} = \bar{\boldsymbol{H}}_\gamma^{-1}[\underline{\boldsymbol{H}}_\gamma\underline{\gamma} + \sigma^{-2}\tilde{\boldsymbol{W}}'\boldsymbol{Q}^{-1}\tilde{\boldsymbol{y}}]$. Draws from this distribution are straightforward.

The conditional posterior density kernel for σ^2 is the product of Equations (A.2, A.9, and A.13). This kernel corresponds to the conditional posterior distribution

$$\left[\underline{s}^2 + \alpha'\underline{\boldsymbol{H}}_\alpha\alpha + \sum_{t=1}^{T}\sigma_{(t)}^{-2}\left(\tilde{y}_t - \alpha'\tilde{\boldsymbol{z}}_t - \beta'\boldsymbol{x}_t\right)^2\right]/\sigma^2 \sim \chi^2(\underline{\nu}+m+T),$$

for σ^2, from which simulation is simple.

In the normal model $\sigma_{(t)}^2 \equiv 1$, and in the normal mixture model $\sigma_{(t)}^2 = \sigma_{s(t)}^2$. In the Student-$t$ model, the kernel for $\sigma_{(t)}^2$ $(t = 1, \ldots, T)$ is the product of Equations (A.2 and A.4). Thus in this model the $\sigma_{(t)}^2$ are conditionally independent, with

$$\left[\lambda + \sigma^{-2}\left(\tilde{y}_t - \beta'\boldsymbol{x}_t\right)^2\right]/\sigma_{(t)}^2 \sim \chi^2(\lambda+1).$$

When the disturbance distribution is Student-t, the conditional posterior density kernel for λ is given by the product of Equations (A.4 and A.10),

$$k(\lambda) = \left[2^{\lambda/2}\Gamma(\lambda/2)\right]^{-T} \lambda^{T\lambda/2} \left[\prod_{t=1}^{T} \sigma^2_{(t)}\right]^{-(\lambda+2)/2} \exp\left\{\left[\underline{\lambda}^{-1} + \left(\tfrac{1}{2}\right)\sum_{t=1}^{T} \sigma^{-2}_{(t)}\right]\lambda\right\}. \tag{A.18}$$

This is a proper density kernel, but it is not the kernel of any conventional p.d.f. We therefore incorporate a Metropolis step within the Gibbs sampling algorithm, as described in Section 2.6. The candidate density $q(\lambda)$ is that of a univariate normal distribution, with mean at the maximum $\hat{\lambda}$ of $k(\lambda)$, and precision equal to $-d^2 \log k(\lambda)/d\lambda^2|_{\lambda=\hat{\lambda}}$. A draw λ^* is taken from this normal distribution at each step m of the MCMC algorithm. With probability

$$\min\left\{\frac{k(\lambda^*)/q(\lambda^*)}{k\left(\lambda^{(m-1)}\right)/q\left(\lambda^{(m-1)}\right)}, 1\right\}, \tag{A.19}$$

$\lambda^{(m)} = \lambda^*$; otherwise $\lambda^{(m)} = \lambda^{(m-1)}$.

In the normal mixture model the conditional posterior density kernel for the state assignments $s(t)$ ($t = 1, \ldots, T$) is the product of Equations (A.5 and A.6) taken over $t = 1, \ldots, T$. Thus the $s(t)$ are conditionally independent, with

$$P[s(t) = j] \propto p_j \sigma_j^{-1} \exp\left[-\left(\tilde{y}_t - \alpha_j - \beta' \boldsymbol{x}_t\right)^2 / 2\sigma^2\sigma_j^2\right] \quad (j = 1, \ldots, m).$$

Draws from these multinomial distributions are trivial. Following these draws, in $\tilde{\boldsymbol{Z}} = [\tilde{z}_{tj}]$ set $\tilde{z}_{tj} = \delta_{s(t),j}$.

In the normal mixture model the conditional posterior density kernel for $\boldsymbol{p}$ is the product of Equations (A.5 and A.11), $\prod_{j=1}^{m} p_j^{\underline{r}_j + T_j - 1}$, where T_j is the number of observations assigned to state j, $T_j = \sum_{t=1}^{T} \delta_{s(t),j}$. Thus the conditional posterior distribution of $\boldsymbol{p}$ is Dirichlet with parameters $\underline{r}_j + T_j$ ($j = 1, \ldots, m$). It is straightforward to draw from this distribution; see Johnson and Kotz (1972, Section 40.5).

The conditional posterior density kernel of $(\sigma_1^2, \ldots, \sigma_m^2)$ in the normal mixture model is the product of Equation (A.6) taken over $t = 1, \ldots, T$ and Equation (A.12), which implies that subject to the restriction $\sigma_1^2 > \cdots > \sigma_m^2$, the σ_j^2 are independent with

$$\left[\underline{s}_j^2 + \sum_{t=1}^{T} \delta_{s(t),j}\left(\tilde{y}_t - \alpha' \tilde{\boldsymbol{z}}_t - \beta' \boldsymbol{x}_t\right)^2\right] / \sigma_j^2 \sim \chi^2\left(\underline{\nu}_j + T_j\right) \quad (j = 1, \ldots, m). \tag{A.20}$$

The ordering restriction is enforced by drawing the σ_j^2 in succession from Equation (A.20). At each step in this succession, if the candidate draw conforms with the ordering it is accepted; if not, the current value of σ_j^2 is retained. Each draw in this succession is thus a Metropolis within Gibbs step.

The latent variables $\tilde{y}_t$ $(t = 1, \ldots, T)$ appear in the components (A.2 and A.7) of the likelihood function. Thus they are conditionally independent,

$$\tilde{y}_t \sim N\left(\alpha'\mathbf{z}_t + \beta\mathbf{x}_t, \sigma^2\sigma_j^2\right) \quad \text{subject to} \quad \tilde{y}_t \in [c_t, d_t].$$

Marginal likelihoods. The Gelfand–Dey harmonic mean algorithm for approximation of the marginal likelihood requires that the prior and data densities be evaluated at each iteration used in the marginal likelihood approximation. The efficiency of the algorithm can be increased by analytical integration of groups or parameters wherever possible. In the ULLM, it is easy to integrate latent variables prior to evaluation of the prior and data densities. When this is done the Gelfand–Dey algorithm is computationally quite efficient.

For normally distributed disturbances the prior density is the product of Equations (A.8 and A.9), and the only latent variables are $\tilde{y}_t$ $(t = 1, \ldots, T)$. Integrating over the $\tilde{y}_t$, the data density is $\prod_{t=1}^{T} p\left(y_t|\mathbf{x}_t, \beta, \sigma^2\right)$. If $c_t = d_t$ then

$$p\left(y_t|\mathbf{x}_t, \beta, \sigma^2\right) = \sigma^{-1}\phi\left[\left(y_t - \beta'\mathbf{x}_t\right)/\sigma\right].$$

If $c_t < d_t$,

$$p\left(y_t|\mathbf{x}_t, \beta, \sigma^2\right) = \Phi\left[\left(d_t - \beta'\mathbf{x}_t\right)/\sigma\right] - \Phi\left[\left(c_t - \beta'\mathbf{x}_t\right)/\sigma\right].$$

Observe that the last expression integrates the latent $\tilde{y}_t$ analytically from the vector of unknown parameters and latent variables.

The MCMC algorithm for the ULLM in the case of Student-t disturbances employs the auxiliary latent variables $\sigma_{(t)}^2$ $(t = 1, \ldots, T)$. These could be regarded as part of the parameter vector, but it is more efficient to carry out the integration over the $\sigma_{(t)}^2$ analytically, just as is done for the $\tilde{y}_t$ when $c_t < d_t$. This yields $\sigma^{-1}t\left[(y_t - \beta'\mathbf{x}_t)/\sigma;\ \lambda\right]$ for the data density when $c_t = d_t$, and $T\left[(d_t - \beta'\mathbf{x}_t)/\sigma;\ \lambda\right] - T\left[(c_t - \beta'\mathbf{x}_t)/\sigma;\ \lambda\right]$ when $c_t < d_t$. The prior density is the product of Equations (A.8, A.9 and A.10).

For the normal mixture model it is efficient to integrate analytically across the latent states $s(t)$ $(t = 1, \ldots, T)$. When this is done, the data density at observation t is

$$\sigma^{-1}\sum_{j=1}^{m} p_j\sigma_j^{-1}\phi\left[\left(y_t - \alpha_j - \beta'\mathbf{x}_t\right)/\sigma\sigma_j\right],$$

if $c_t = d_t$, and

$$\sum_{j=1}^{m} p_j\left\{\Phi\left[\left(d_t - \alpha_j - \beta'\mathbf{x}_t\right)/\sigma\sigma_j\right] - \Phi\left[\left(c_t - \alpha_j - \beta'\mathbf{x}_t\right)/\sigma\sigma_j\right]\right\},$$

if $c_t < d_t$.

Appendix B. The full multivariate latent linear model

Core model. The core multivariate latent linear model (MLLM) is

$$\tilde{\boldsymbol{y}}_t = \boldsymbol{A}'\tilde{\boldsymbol{z}}_t + \boldsymbol{B}'\boldsymbol{x}_t + \varepsilon_t \quad (t = 1, \ldots, T). \tag{B.1}$$

The $k \times 1$ vector $\boldsymbol{x}_t$ and $m \times 1$ vector $\tilde{\boldsymbol{z}}_t$ consist of observed covariates and latent variables, respectively, just as in the ULLM; $k > 0$ and $m \geqslant 0$. Take $\tilde{\boldsymbol{Y}}' = [\tilde{\boldsymbol{y}}_1, \ldots, \tilde{\boldsymbol{y}}_T]$, $\tilde{\boldsymbol{Z}}' = [\tilde{\boldsymbol{z}}_1, \ldots, \tilde{\boldsymbol{z}}_T]$, $\boldsymbol{X}' = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T]$, and $\boldsymbol{E}' = [\varepsilon_1, \ldots, \varepsilon_T]$. Then Equation (B.1) may be expressed

$$\tilde{\boldsymbol{Y}} = \tilde{\boldsymbol{Z}}\boldsymbol{A} + \boldsymbol{X}\boldsymbol{B} + \boldsymbol{E}. \tag{B.2}$$

Defining $\tilde{\boldsymbol{y}} = \mathrm{vec}(\tilde{\boldsymbol{Y}})$, $\alpha = \mathrm{vec}(\boldsymbol{A})$, $\beta = \mathrm{vec}(\boldsymbol{B})$ and $\varepsilon = \mathrm{vec}(\boldsymbol{E})$, it may also be expressed[36]

$$\tilde{\boldsymbol{y}} = \left(\boldsymbol{I}_p \otimes \tilde{\boldsymbol{Z}}\right) \alpha + \left(\boldsymbol{I}_p \otimes \boldsymbol{X}\right) \beta + \varepsilon.$$

The outcome vector $\tilde{\boldsymbol{y}}_t$ has p elements. Some (or all) of these elements may be observed, and some (or all) of them may be latent subject to linear restrictions discussed subsequently. The $m \times p$ matrix of parameters $\boldsymbol{A}$ and the $k \times p$ matrix of parameters $\boldsymbol{B}$ are unknown.

Degeneracy in ε_t arises if there is a $p \times g$ matrix $\boldsymbol{G}_0$ such that $\boldsymbol{G}_0'\tilde{\boldsymbol{y}}_t = \boldsymbol{g}_0 \ \forall\ t$, and hence, $\boldsymbol{G}_0'\varepsilon_t \equiv \boldsymbol{0} \ \forall\ t$. Take $q = p - g$ and let $\boldsymbol{G}$ be any $p \times q$ orthonormal matrix of rank q such that $\boldsymbol{G}_0'\boldsymbol{G} = \boldsymbol{0}$. The nondegenerate components of ε_t are $\varepsilon_t^* = \boldsymbol{G}'\varepsilon_t$, and the conditionally nondegenerate components of $\tilde{\boldsymbol{y}}_t$ are $\tilde{\boldsymbol{y}}_t^* = \boldsymbol{G}'\tilde{\boldsymbol{y}}_t$. The core distributional assumption may be stated $\varepsilon_t^* = \eta_t\sigma_{(t)}$, where $\sigma_{(t)}$ is a strictly positive, i.i.d. random variable conditional on $\boldsymbol{X}$, and

$$\eta_t | \left(\sigma_{(1)}, \ldots, \sigma_{(T)}, \boldsymbol{X}\right) \overset{\text{i.i.d.}}{\sim} N(0, \Sigma).$$

Hence

$$p\left(\tilde{\boldsymbol{Y}} | \boldsymbol{X}, \tilde{\boldsymbol{Z}}, \{\sigma_{(t)}^2\}_{t=1}^T\right) = (2\pi)^{-q/2} |\Sigma|^{-T/2} \left[\prod_{t=1}^T \sigma_{(t)}^2\right]^{-q/2} \cdot \exp\left[-\sum_{t=1}^T \left(\tilde{\boldsymbol{y}}_t - \boldsymbol{A}'\tilde{\boldsymbol{z}}_t - \boldsymbol{B}'\boldsymbol{x}_t\right)' \boldsymbol{G}\Sigma^{-1}\boldsymbol{G}' \left(\tilde{\boldsymbol{y}}_t - \boldsymbol{A}'\tilde{\boldsymbol{z}}_t - \boldsymbol{B}'\boldsymbol{x}_t\right) / 2\sigma_{(t)}^2\right] \tag{B.3}$$

$$= (2\pi)^{-q/2} |\Sigma|^{-T/2} \left[\prod_{t=1}^T \sigma_{(t)}^2\right]^{-q/2} \exp\left\{-(1/2)\mathrm{tr}\left[\left(\sum_{t=1}^T \varepsilon_t^* \varepsilon_t^{*\prime}\right) \Sigma\right]\right\}, \tag{B.4}$$

it being understood that the support of the density is limited to $\tilde{\boldsymbol{Y}} : \tilde{\boldsymbol{Y}}\boldsymbol{G}_0 = \boldsymbol{g}_0'$.

[36] For any $k \times m$ matrix $\boldsymbol{A}$ and $m \times n$ matrix $\boldsymbol{B}$, $\mathrm{vec}(\boldsymbol{AB}) = (\boldsymbol{I}_n \otimes \boldsymbol{A})\ \mathrm{vec}(\boldsymbol{B}) = (\boldsymbol{B}' \otimes \boldsymbol{I}_k)\ \mathrm{vec}(\boldsymbol{A})$. These facts are used often in this section.

As in the ULLM, the latent vector $\tilde{\boldsymbol{z}}_t$ arises in the normal mixture distribution of the disturbances. The vector $\tilde{\boldsymbol{z}}_t$ is determined by the state assignment process, which is multinomial and independent of η_t. Hence $\tilde{\boldsymbol{Y}}\boldsymbol{G}_0 = \boldsymbol{g}_0'$ implies $\boldsymbol{A}\boldsymbol{G}_0 = \boldsymbol{0}$, and if $\boldsymbol{A}^* = \boldsymbol{A}\boldsymbol{G}$ is known then so is $\boldsymbol{A}$. For subsequent purposes it is useful to take $\alpha^* = \text{vec}(\boldsymbol{A}^*) = \text{vec}(\boldsymbol{A}\boldsymbol{G}) = (\boldsymbol{G}' \otimes \boldsymbol{I}_m)\alpha$ as the vector of unknown parameters in $\boldsymbol{A}$.

There will be similar restrictions on $\boldsymbol{B}$, centered on the rows of $\boldsymbol{B}$ corresponding to the intercept term in $\boldsymbol{x}_t$. The implications for restrictions on $\boldsymbol{B}$ can be considerably richer and more varied than for those on $\boldsymbol{A}$, however. For example, in a multinomial probit model some covariates may be specific to individuals, others to choices. To incorporate these and perhaps other restrictions, define $\beta = \text{vec}(\boldsymbol{B})$ and take the linear restrictions to be of the general form $\boldsymbol{R}_1\beta = \boldsymbol{r}_1$, where $\boldsymbol{R}_1$ is an $\ell_1 \times pk$ matrix of rank ℓ_1. Define the $pk \times pk$ matrix $\boldsymbol{P}' = [\boldsymbol{P}_1' \vdots \boldsymbol{P}_2']$ in which $\boldsymbol{P}_1 = \boldsymbol{R}_1$, $\boldsymbol{P}_1\boldsymbol{P}_2' = \boldsymbol{0}$, and $\boldsymbol{P}_2\boldsymbol{P}_2' = \boldsymbol{I}_{pk-\ell_1}$. Then $\boldsymbol{P}^{-1} = [\boldsymbol{P}_1'(\boldsymbol{P}_1\boldsymbol{P}_1')^{-1} \vdots \boldsymbol{P}_2']$. Transform β to $\beta^* = \boldsymbol{P}\beta$. The first ℓ_1 elements of β^* are $\beta_1^* = \boldsymbol{r}_1$, and the remaining $\ell_2 = pk - \ell_1$ elements are the unknown parameters $\beta_2^* = \boldsymbol{P}_2\beta$.

Concentrating on the nondegenerate component of the equation system (B.2), let $\tilde{\boldsymbol{Y}}^* = \tilde{\boldsymbol{Y}}\boldsymbol{G}$ and $\boldsymbol{E}^* = \boldsymbol{E}\boldsymbol{G}$; $\text{vec}(\tilde{\boldsymbol{Y}}^*) = \tilde{\boldsymbol{y}}^*$ and $\text{vec}(\boldsymbol{E}^*) = \varepsilon^*$. From $\tilde{\boldsymbol{Y}}^* = \tilde{\boldsymbol{Z}}\boldsymbol{A}\boldsymbol{G} + \boldsymbol{X}\boldsymbol{B}\boldsymbol{G} + \boldsymbol{E}^*$,

$$\tilde{\boldsymbol{y}}^* = \left(\boldsymbol{I}_q \otimes \tilde{\boldsymbol{Z}}\right)\, \alpha^* + \left(\boldsymbol{I}_q \otimes \boldsymbol{X}\right)\, \text{vec}(\boldsymbol{B}\boldsymbol{G}) + \varepsilon^*. \tag{B.5}$$

Incorporating the definitions of α^* and β^* Equation (B.5) becomes

$$\tilde{\boldsymbol{y}}^* = \left(\boldsymbol{I}_q \otimes \tilde{\boldsymbol{Z}}\right)\, \alpha^* + \left(\boldsymbol{I}_q \otimes \boldsymbol{X}\right)\, \left(\boldsymbol{G}' \otimes \boldsymbol{I}_k\right)\, \boldsymbol{P}^{-1}\beta^* + \varepsilon^*.$$

Defining $\underset{qT \times pk}{\boldsymbol{W}} = \begin{bmatrix} \underset{qT \times \ell_1}{\boldsymbol{W}_1} & \vdots & \underset{qT \times \ell_2}{\boldsymbol{W}_2} \end{bmatrix} = \left(\boldsymbol{I}_q \otimes \boldsymbol{X}\right)(\boldsymbol{G}' \otimes \boldsymbol{I}_k)\, \boldsymbol{P}^{-1}$, we have

$$\tilde{\boldsymbol{y}}^* - \boldsymbol{W}_1\boldsymbol{r}_1 = \left(\boldsymbol{I}_q \otimes \tilde{\boldsymbol{Z}}\right)\, \alpha^* + \boldsymbol{W}_2\,\beta_2^* + \varepsilon^*.$$

Define $\boldsymbol{Q} = \text{diag}[\sigma_{(1)}^2, \ldots, \sigma_{(T)}^2]$ just as in the ULLM. Since $\varepsilon^* | \boldsymbol{Q} \sim N(\boldsymbol{0}, \Sigma \otimes \boldsymbol{Q})$,

$$\begin{aligned} p\left(\tilde{\boldsymbol{y}} | \boldsymbol{X}, \tilde{\boldsymbol{Z}}, \boldsymbol{Q}\right) = {} & (2\pi)^{-qT/2}\, |\Sigma|^{-T/2}\, |\boldsymbol{Q}|^{-q/2} \\ & \cdot \exp\left\{-\tfrac{1}{2}\left[\left(\boldsymbol{G}' \otimes \boldsymbol{I}_T\right)\tilde{\boldsymbol{y}} - \boldsymbol{W}_1\boldsymbol{r}_1 - \left(\boldsymbol{I}_q \otimes \tilde{\boldsymbol{Z}}\right)\, \alpha^* - \boldsymbol{W}_2\,\beta_2^*\right]'\left(\Sigma^{-1} \otimes \boldsymbol{Q}^{-1}\right)\right. \\ & \left. \cdot \left[\left(\boldsymbol{G}' \otimes \boldsymbol{I}_T\right)\tilde{\boldsymbol{y}} - \boldsymbol{W}_1\boldsymbol{r}_1 - \left(\boldsymbol{I}_q \otimes \tilde{\boldsymbol{Z}}\right)\, \alpha^* - \boldsymbol{W}_2\,\beta_2^*\right]\right\}, \end{aligned} \tag{B.6}$$

it being understood that the support of Equation (B.6) is $\tilde{\boldsymbol{y}} : (\boldsymbol{G}'_0 \otimes \boldsymbol{I}_T)\tilde{\boldsymbol{y}} = \boldsymbol{0}$.

Distributional assumptions. There is a rich variety of distributions within this framework. We consider in detail the same three treated in the ULLM: normal, Student-*t*, and normal mixture.

When $m = 0$ and $\sigma_{(t)}^2 \equiv 1$, $\varepsilon_t^* \overset{\text{i.i.d.}}{\sim} N(\mathbf{0}, \Sigma)$. When $m = 0$ and $\lambda\sigma_{(t)}^{-2} \overset{\text{i.i.d.}}{\sim} \chi^2(\lambda)$, then $p(\sigma_{(t)}^2|\lambda, \boldsymbol{X})$ is again given by Equation (A.4). The distribution of ε_t^* is multivariate Student-t with scale matrix Σ and λ degrees of freedom: $\varepsilon_t^*|(\lambda, \boldsymbol{X}) \overset{\text{i.i.d.}}{\sim} t(\mathbf{0}, \Sigma; \lambda)$.

The normal mixture distribution is similar to that in the ULLM. There are $m \geqslant 2$ latent states. States are selected independently over the observations, with state j having probability p_j as indicated in Equation (A.5). Conditional on state j, $\varepsilon_t^* = \boldsymbol{G}'\varepsilon_t$ has mean given by α_j^* (the jth column of $\boldsymbol{G}'\boldsymbol{A}'$) and variance $\sigma_j^2\Sigma$. Thus

$$p\left(\tilde{\boldsymbol{y}}_t|\boldsymbol{x}_t, s(t) = j\right) = (2\pi)^{-q/2}\sigma_j^{-q}\left|\Sigma\right|^{-1/2} \cdot \exp\left[-\tfrac{1}{2}\left(\boldsymbol{G}'\tilde{\boldsymbol{y}}_t - \alpha_j^* - \boldsymbol{G}'\boldsymbol{B}'\boldsymbol{x}_t\right)'\Sigma^{-1}\left(\boldsymbol{G}'\tilde{\boldsymbol{y}}_t - \alpha_j^* - \boldsymbol{G}'\boldsymbol{B}'\boldsymbol{x}_t\right)\right]. \tag{B.7}$$

Prior distributions. For each group of parameters in the MLLM, there is a benchmark proper prior, much as in the ULLM. The MLLM priors are similar, but differ in some details because of complications in the MLLM not present in the ULLM.

To complement the set of ℓ_1 linear restrictions $\boldsymbol{R}_1\beta = \boldsymbol{r}_1$ imposed on $\beta = \mathrm{vec}(\boldsymbol{B})$, take the remaining prior distribution for β in the form $\boldsymbol{R}_2\beta \sim N(\boldsymbol{r}_2, \boldsymbol{V}_2)$. The matrix $\left[\boldsymbol{R}_1' : \boldsymbol{R}_2'\right]$ must be of rank pk, so $\boldsymbol{R}_2 : \ell_2 \times pk$ must have $\ell_2 \geqslant pk - \ell_1$. Referring to the definition of $\beta^* = \boldsymbol{P}\beta$ above, it follows that $\beta_1^* = \boldsymbol{r}_1$ and $\beta_2^* \sim N\left(\underline{\beta}_2^*, \underline{\boldsymbol{H}}_{\beta_2^*}^{-1}\right)$,

$$p\left(\beta_2^*\right) = (2\pi)^{-\ell_2/2}\left|\underline{\boldsymbol{H}}_{\beta_2^*}\right|^{1/2}\exp\left[-\tfrac{1}{2}\left(\beta_2^* - \underline{\beta}_2^*\right)'\underline{\boldsymbol{H}}_{\beta_2^*}\left(\beta_2^* - \underline{\beta}_2^*\right)\right], \tag{B.8}$$

with

$$\underline{\boldsymbol{H}}_{\beta_2^*} = \boldsymbol{P}_2\boldsymbol{R}_2'\boldsymbol{V}_2^{-1}\boldsymbol{R}_2\boldsymbol{P}_2' \text{ and } \underline{\beta}_2^* = \underline{\boldsymbol{H}}_{\beta_2^*}^{-1}\boldsymbol{P}_2\boldsymbol{R}_2'\boldsymbol{V}_2^{-1}\left[\boldsymbol{r}_2 - \boldsymbol{R}_2\boldsymbol{P}_1'\left(\boldsymbol{P}_1\boldsymbol{P}_1'\right)^{-1}\boldsymbol{r}_1\right]. \tag{B.9}$$

Clearly the representation $\boldsymbol{R}_2\beta \sim N(\boldsymbol{r}_2, \boldsymbol{V}_2)$ is not unique: all that matters is $\underline{\boldsymbol{H}}_{\beta_2^*}$ and $\underline{\beta}_2^*$. However, it is often convenient to represent prior information about β in individual, independent components, so that $\boldsymbol{V}_2$ is diagonal.

We employ a conventional inverted Wishart prior distribution, $\Sigma^{*\,-1} \sim W(\underline{\boldsymbol{S}}^{-1}, \underline{\nu})$, for Σ^*:

$$p\left(\Sigma^*\right) = 2^{-\underline{\nu}q/2}\pi^{-q(q-1)/4}\prod_{j=1}^{q}\Gamma\left[(\underline{\nu} + 1 - j)/2\right]^{-1} \cdot \left|\underline{\boldsymbol{S}}^*\right|^{\underline{\nu}/2}\left|\Sigma^*\right|^{-(\underline{\nu}+q+1)/2}\exp\left[-\tfrac{1}{2}\mathrm{tr}\underline{\boldsymbol{S}}^*\Sigma^{*\,-1}\right]. \tag{B.10}$$

If all elements of $\tilde{\boldsymbol{y}}_t^*$ are observed, then the prior distribution is the conditionally conjugate one for the variance matrix in a system of seemingly unrelated regressions; see Chib and Greenberg (1995b) for discussion.

When $q < p$ then interactions between the choice of $\boldsymbol{G}$ and the prior distribution for Σ must be taken into account. Consider the alternative $\tilde{\varepsilon}_t^* = \tilde{\boldsymbol{G}}'\varepsilon_t$ to $\varepsilon_t^* = \boldsymbol{G}'\varepsilon_t$ where $\tilde{\boldsymbol{G}}$ is also orthonormal and $\tilde{\boldsymbol{G}}'\boldsymbol{G}_0 = \boldsymbol{0}$. Then $\mathrm{var}(\tilde{\varepsilon}_t^*) = \tilde{\Sigma} = \tilde{\boldsymbol{G}}'\boldsymbol{G}\,\Sigma\,\boldsymbol{G}'\tilde{\boldsymbol{G}}$. The $q \times q$ matrix $\tilde{\boldsymbol{G}}'\boldsymbol{G}$ is orthonormal of rank q, so if $\Sigma^* \sim IW(\underline{s}^2\boldsymbol{I}_q, \underline{\nu})$, then $\tilde{\Sigma}^* \sim IW(\underline{s}^2\boldsymbol{I}_q, \underline{\nu})$, as well. Since $\mathrm{tr}\tilde{\Sigma}^* = \mathrm{tr}\Sigma^*$ it follows that in the multinomial probit case the prior distribution $\Sigma^* \sim IW(\underline{s}^2\boldsymbol{I}_q, \underline{\nu})$ is invariant to the choice of $\boldsymbol{G}$. Given the arbitrary scaling in the model, it is convenient to take $\underline{s}^2 = \underline{\nu}$. Note that in the multinomial probit model $\mathrm{var}(\varepsilon_t^*) \propto \boldsymbol{I}_q$ corresponds to a construction in which $\varepsilon_t^0 \sim N(\boldsymbol{0}, \boldsymbol{I}_p)$, $\varepsilon_t = \varepsilon_t^0 - \boldsymbol{e}_p p^{-1}\boldsymbol{e}_p'\varepsilon_t$, for any choice of $\boldsymbol{G}$. Thus, $\Sigma^* \sim IW(\underline{\nu}\boldsymbol{I}_q, \underline{\nu})$ is an attractive benchmark prior for the multinomial probit model: larger values of $\underline{\nu}$ correspond to prior beliefs that choice-specific random components of utility are more nearly independent[37].

Given our construction of $\boldsymbol{G}$ in the more general case in which one or more proper subsets of the p equations corresponds to a set of exhaustive discrete choices, these considerations apply to each set considered separately. The rows and columns of $\underline{\boldsymbol{S}}$ corresponding to each set are of the form $\underline{\nu}\boldsymbol{I}$, and $\underline{\boldsymbol{S}}$ has a block diagonal structure reflecting the sets of exhaustive discrete choices.

This completes the prior distribution when the disturbances are normal, because in that case $\beta = \mathrm{vec}(\boldsymbol{B})$ and Σ are the only parameters in the model. The Student-t MLLM has one additional parameter, λ. The prior for λ is the same exponential distribution used in the ULLM (Equation A.10). The normal mixture model has three additional parameters: the vector of state probabilities $\boldsymbol{p}$, the state variances σ_j^2 $(j = 1, \ldots, m)$, and the matrix of state mean vectors $\boldsymbol{A}$. The first two of these play the same role in the MLLM and in the ULLM, and their prior distributions are of the same form: Dirichlet for $\boldsymbol{p}$ (Equation A.11), and independent inverted gamma for the σ_j^2 subject to the ordering restrictions (A.12).

The rows of the $m \times p$ matrix $\boldsymbol{A}$ are the means of the disturbance vectors in each of the m states. Restrictions of the form $\boldsymbol{G}_0'\tilde{\boldsymbol{y}}_t = \boldsymbol{g}_0$ imply restrictions on means and (more generally) on the coefficients of covariates. Earlier in this section we have adopted the convention that the restrictions $\boldsymbol{G}_0'\boldsymbol{e}_p = \boldsymbol{g}_0$ are enforced explicitly through components of the linear restrictions $\boldsymbol{R}_1\,\beta = \boldsymbol{r}_1$, leaving the restrictions $\boldsymbol{A}\,\boldsymbol{G}_0 = \boldsymbol{0}$, and therefore $\alpha^* = \mathrm{vec}(\boldsymbol{A}\,\boldsymbol{G}) = (\boldsymbol{G}' \otimes \boldsymbol{I}_m)\,\alpha$ as the vector of unknown parameters in $\boldsymbol{A}$. For the same reasons discussed in the ULLM it is useful to condition on Σ, and take the mean to be $\boldsymbol{0}$ and the variance proportional to Σ: $\alpha^* \sim N(\boldsymbol{0}, \Sigma \otimes \underline{\boldsymbol{H}}_{\alpha^*}^{-1})$:

$$p(\alpha^*) = (2\pi)^{-mq/2}\,|\Sigma|^{-m/2}\,|\underline{\boldsymbol{H}}_{\alpha^*}|^{q/2}\exp\left[\left(-\tfrac{1}{2}\right)\alpha^{*\prime}\left(\Sigma^{-1} \otimes \underline{\boldsymbol{H}}_{\alpha^*}\right)\,\alpha^*\right]. \qquad \text{(B.11)}$$

The form $\Sigma \otimes \underline{\boldsymbol{H}}_{\alpha}^{-1}$ is consistent with the prior distribution employed in the ULLM, leads to a tractable conditional posterior distribution for Σ subsequently, and is

[37] In the conventional treatment [see Danise (1985), Bunch (1991) or Geweke, Keane and Runkle (1994)], an arbitrary "last choice" equation is subtracted from the others. This yields a matrix $\boldsymbol{G}$ with $\boldsymbol{G}'\boldsymbol{G}_0 = \boldsymbol{0}$, but since the columns of $\boldsymbol{G}$ are not orthonormal, prior distributions must change with the choice of which choice is "last".

invariant to the choice of the orthonormal matrix $\boldsymbol{G}$. In all of the work described here, $\underline{\boldsymbol{H}}_{\alpha^*} = \underline{\boldsymbol{h}}_{\alpha^*}\boldsymbol{I}_m$. Prior beliefs about unimodality are reflected in the choice of $\underline{\boldsymbol{h}}_{\alpha^*}$ in the MLLM just as they are in the choice of $\underline{\boldsymbol{h}}_{\alpha}$ in the ULLM.

Inference in the MLLM. There are eight groups of parameters or latent variables in the model: (α^*, β_2^*); Σ^*; $\sigma_{(t)}^2$ $(t = 1, \ldots, T)$; λ; $s(t)$ $(t = 1, \ldots, T)$ and $\tilde{\boldsymbol{Z}}$; $\boldsymbol{p}$; σ_j^2 $(j = 1, \ldots, m)$ and $\tilde{\boldsymbol{y}}_t$ $(t = 1, \ldots, T)$. As in the ULLM, not all parameters appear under each distributional assumption. The posterior density kernel is the product of the prior and data densities that apply in the model at hand. As in the ULLM, we construct a MCMC posterior simulator to access the posterior distribution.

For α^* and β_2^* let

$$\gamma^* = \begin{pmatrix} \alpha^* \\ \beta_2^* \end{pmatrix}, \quad \underline{\gamma}^* = \begin{pmatrix} \mathbf{0} \\ \underline{\beta}_2^* \end{pmatrix}, \quad \text{and} \quad \underline{\boldsymbol{H}}_{\gamma^*} = \begin{bmatrix} \Sigma^{-1} \otimes \underline{\boldsymbol{H}}_{\alpha^*} & \mathbf{0} \\ \mathbf{0} & \underline{\boldsymbol{H}}_{\beta_2^*} \end{bmatrix},$$

with $\underline{\beta}_2^*$ and $\underline{\boldsymbol{H}}_{\beta_2^*}$ defined in Equation (B.9). The conditional posterior density kernel is the product of Equations (B.6, B.8 and B.11), implying the conditional distribution is $\gamma^* \sim N(\bar{\gamma}^*, \bar{\boldsymbol{H}}_{\gamma^*})$ with

$$\bar{\boldsymbol{H}}_{\gamma^*} = \begin{bmatrix} \Sigma^{-1} \otimes \left(\underline{\boldsymbol{H}}_{\alpha^*} + \tilde{\boldsymbol{Z}}'\boldsymbol{Q}^{-1}\tilde{\boldsymbol{Z}}\right) & \left(\Sigma^{-1} \otimes \tilde{\boldsymbol{Z}}'\boldsymbol{Q}^{-1}\right)\boldsymbol{W}_2 \\ \boldsymbol{W}'_2\left(\Sigma^{-1} \otimes \boldsymbol{Q}^{-1}\tilde{\boldsymbol{Z}}\right) & \underline{\boldsymbol{H}}_{\beta_2^*} + \boldsymbol{W}'_2\left(\Sigma^{-1} \otimes \boldsymbol{Q}^{-1}\right)\boldsymbol{W}_2 \end{bmatrix},$$

$$\bar{\gamma}^* = \bar{\boldsymbol{H}}_{\gamma^*}^{-1}\left\{\begin{pmatrix} \Sigma^{-1} \otimes \underline{\boldsymbol{H}}_{\alpha^*} \\ \underline{\boldsymbol{H}}_{\beta_2^*}\underline{\beta}_2^* \end{pmatrix} + \begin{pmatrix} \boldsymbol{I}_q \otimes \tilde{\boldsymbol{Z}}' \\ \boldsymbol{W}'_2 \end{pmatrix}\left(\Sigma^{-1} \otimes \boldsymbol{Q}^{-1}\right)\left[\left(\boldsymbol{G}' \otimes \boldsymbol{I}_T\right)\tilde{\boldsymbol{y}} - \boldsymbol{W}_1\boldsymbol{r}_1\right]\right\}.$$

For computational purposes it is useful to write $\bar{\gamma}^* = \bar{\boldsymbol{H}}_{\gamma^*}^{-1}\bar{\boldsymbol{c}}_{\gamma^*}$, define

$$\begin{bmatrix} \tilde{\boldsymbol{W}}_{11} & \tilde{\boldsymbol{W}}_{12} & \tilde{\boldsymbol{W}}_{13} \\ \tilde{\boldsymbol{W}}_{21} & \boldsymbol{W}_{22} & \tilde{\boldsymbol{W}}_{23} \end{bmatrix} = \begin{bmatrix} \tilde{\boldsymbol{Z}}'\boldsymbol{Q}^{-1}\tilde{\boldsymbol{Z}} & \tilde{\boldsymbol{Z}}'\boldsymbol{Q}^{-1}\boldsymbol{X} & \tilde{\boldsymbol{Z}}'\boldsymbol{Q}^{-1}\tilde{\boldsymbol{Z}}\tilde{\boldsymbol{Y}} \\ \boldsymbol{X}'\boldsymbol{Q}^{-1}\tilde{\boldsymbol{Z}} & \boldsymbol{X}'\boldsymbol{Q}^{-1}\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{Q}^{-1}\tilde{\boldsymbol{Z}}\tilde{\boldsymbol{Y}} \end{bmatrix},$$

and then derive

$$\left[\bar{\boldsymbol{H}}_{\gamma^*} \vdots \bar{\boldsymbol{c}}_{\gamma^*}\right] = \begin{bmatrix} \boldsymbol{I}_{qm} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{P}^{2\prime}\left(\boldsymbol{G} \otimes \boldsymbol{I}_k\right) \end{bmatrix} \cdot \begin{bmatrix} \Sigma^{-1} \otimes \tilde{\boldsymbol{W}}_{11} & \Sigma^{-1} \otimes \tilde{\boldsymbol{W}}_{12} & \Sigma^{-1} \otimes \tilde{\boldsymbol{W}}_{13} \\ \Sigma^{-1} \otimes \tilde{\boldsymbol{W}}_{21} & \Sigma^{-1} \otimes \boldsymbol{W}_{22} & \Sigma^{-1} \otimes \tilde{\boldsymbol{W}}_{23} \end{bmatrix}$$
$$\cdot \begin{bmatrix} \boldsymbol{I}_{qm} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \left(\boldsymbol{G}' \otimes \boldsymbol{I}_k\right)\boldsymbol{P}^2 & -\left(\boldsymbol{G}' \otimes \boldsymbol{I}_k\right)\boldsymbol{P}^1\boldsymbol{r}_1 \\ \mathbf{0} & \mathbf{0} & \operatorname{vec}(\boldsymbol{G}) \end{bmatrix}.$$

There are no unknowns in the first and third matrices on the right side of the last equation, so these matrices only need be computed once, at the start of the MCMC algorithm. New computations each iteration are confined to the second matrix.

The conditional posterior density kernel for Σ^* is the product of the likelihood (B.4) and the prior densities (B.10 and B.11), it being understood that $\Sigma = \Delta(\Sigma^*) \cdot \Sigma^* \cdot \Delta(\Sigma^*)$. Thus, the conditional density kernel is

$$\begin{aligned} h(\Sigma^*) = {} & |\Sigma^*|^{-(\underline{\nu}+q+m+T+1)/2} \, |\Delta(\Sigma^*)|^{-(m+T)} \\ & \cdot \exp\left\langle -\tfrac{1}{2} \mathrm{tr}\left\{ \left[\underline{\boldsymbol{S}}^* + \Delta(\Sigma^*)^{-1} \tilde{\boldsymbol{S}} \Delta(\Sigma^*)^{-1} \right] \Sigma^{*-1} \right\} \right\rangle, \end{aligned} \tag{B.12}$$

where $\tilde{\boldsymbol{S}} = \boldsymbol{A}^{*\prime} \underline{\boldsymbol{H}}_{\alpha^*} \boldsymbol{A}^* + \sum_{t=1}^{T} \varepsilon_t^* \varepsilon_t^{*\prime}$. If there are no exhaustive sets of discrete choices represented in the MLLM then $\Delta(\Sigma^*) = \boldsymbol{I}_q$ and Equation (B.12) is the kernel of an inverted Wishart distribution for Σ^*. If $\Delta(\Sigma^*) \neq \boldsymbol{I}_q$ then Equation (B.12) is finitely integrable in Σ^*, and is therefore a conditional density kernel, but is not of any familiar form. We cope with Equation (B.12) by using a Metropolis within Gibbs step. The candidate distribution is multivariate normal. The mean of the candidate distribution is the vector of nonredundant elements of $\hat{\Sigma}^* = \mathrm{argmax}\, h(\Sigma^*)$, and the precision is the negative of the Hessian of $h(\cdot)$ evaluated at $\hat{\Sigma}^*$[38]. Drawing from the candidate distribution and evaluating $h(\cdot)$ and the multivariate normal kernel at the candidate draw is considerably less time consuming than determining $\hat{\Sigma}^*$. It is therefore not unreasonable to undertake many Metropolis steps for Σ^* at each iteration, drawing candidates and replacing the previous draws or earlier candidates according to the usual Metropolis arithmetic analogous to Equation (A.19). In the work reported in Section 6.2, we have used 400 such iterations. Even with this number, finding $\hat{\Sigma}^*$ and making the draws consumes less time than drawing $\tilde{\boldsymbol{Y}}$ in a typical application with more than a few hundred observations.

As in the ULLM, $\sigma_{(t)}^2 \equiv 1$ when disturbances are normal, and $\sigma_{(t)}^2 = \sigma_{s(t)}^2$ in the normal mixture model. In the Student-t model, Equations (B.3 and A.4) imply the conditional distribution for $\sigma_{(t)}^2$

$$\left(\lambda + \varepsilon_t^{*\prime} \Sigma^{-1} \varepsilon_t^*\right) / \sigma_{(t)}^2 \sim \chi^2(\lambda + q) \quad (t = 1, \ldots, T),$$

and the conditional posterior density kernel for the degrees of freedom parameter λ is again given by Equation (A.18).

The conditional posterior density for the state assignments $s(t)$ $(t = 1, \ldots, T)$ in the normal mixture model is similar to that in the ULLM. From Equations (A.5 and B.7), the $s(t)$ are conditionally independent, with

$$P[s(t) = j] \propto p_j \sigma_j^{-q} \exp\left[-\left(\boldsymbol{G}'\boldsymbol{y}_t - \alpha_j^* - \boldsymbol{G}'\boldsymbol{B}'\boldsymbol{x}_t\right)' \Sigma^{*-1} \left(\boldsymbol{G}'\boldsymbol{y}_t - \alpha_j^* - \boldsymbol{G}'\boldsymbol{B}'\boldsymbol{x}_t\right) / 2\sigma_j^2 \right].$$

Given $s(t) = j$, $\tilde{z}_{tj} = \delta_{s(t),j}$. The conditional posterior distribution for $\boldsymbol{p}$ is again Dirichlet with parameters $(\underline{r}_j + T_j)$, T_j being the number of observations assigned to state j.

[38] Analytic first derivatives of $h(\cdot)$ are straightforward. The maximum of $h(\cdot)$ is found using a quasi-Newton method with a positive definite approximation of the Hessian. The Hessian at $\hat{\Sigma}^*$ is approximated using finite differences of the gradient. See Dennis and Schnabel (1983, Appendix A) and IMSL (1994, Chapter 8).

From Equations (B.7 and A.12), the joint conditional distribution of σ_j^2 $(j = 1, \ldots, m)$ consists of the components

$$\left(\underline{s}_j^2 + \sum_{t=1}^{T} \delta_{j,s(t)} \varepsilon_t^{*\prime} \Sigma^{-1} \varepsilon_t^* \right) / \sigma_j^2 \sim \chi^2 \left(\underline{\nu}_j + qT_j \right) \quad (j = 1, \ldots, m),$$

independent but subject to the ordering restriction $\sigma_1^2 > \cdots > \sigma_m^2$. The ordering restriction is enforced through the same Metropolis rejection procedure used in the ULLM.

In the MLLM, the latent vectors $\tilde{\boldsymbol{y}}_t^* (t = 1, \ldots, T)$ are conditionally independent, with distribution (B.3) restricted by Equation (6.2):

$$\tilde{\boldsymbol{y}}_t^* \sim N \left(\boldsymbol{A}'\boldsymbol{z}_t + \boldsymbol{G}'\boldsymbol{B}'\boldsymbol{x}_t, \sigma_{(t)}^2 \Sigma^* \right) \text{ subject to } \boldsymbol{c}_t^* \leqslant \boldsymbol{F}_t^* \tilde{\boldsymbol{y}}_t^* \leqslant \boldsymbol{d}_t^*,$$

where $\boldsymbol{c}_t^* = \boldsymbol{c}_t - \boldsymbol{F}_t \boldsymbol{G}_0 (\boldsymbol{G}_0' \boldsymbol{G}_0)^{-1} \boldsymbol{g}_0$, $\boldsymbol{d}_t^* = \boldsymbol{d}_t - \boldsymbol{F}_t \boldsymbol{G}_0 (\boldsymbol{G}_0' \boldsymbol{G}_0)^{-1} \boldsymbol{g}_0$ and $\boldsymbol{F}_t^* = \boldsymbol{F}_t \boldsymbol{G}$. The problem of drawing $\tilde{\boldsymbol{y}}_t^*$ can be broken down into drawing successively from the individual components of $\tilde{\boldsymbol{z}}_t^* = \boldsymbol{F}_t^* \tilde{\boldsymbol{y}}_t^*$, each of which is conditionally normal. (Of course, if $c_{tj} = d_{tj}$, then $\tilde{y}_{tj}^*$ is fixed.) Details are given in Geweke (1991).

The continuous state Markov chain defined by this algorithm is ergodic, for the same reasons the ULLM Markov chain is ergodic. Except for the parameters σ_j^2 $(j = 1, \ldots, m)$ in the normal mixture model the transition probability from any point to any subset of positive posterior probability in a single iteration is positive. This is so even for $\tilde{\boldsymbol{y}}_t^*$, because draws are made successively for the $\tilde{\boldsymbol{z}}_t^*$, and for each of the latent variables in $\tilde{\boldsymbol{z}}_t^*$ the support of the conditional coincides with the support of the marginal posterior distribution.

Marginal likelihoods. The algorithm used to evaluate the marginal likelihood in the ULLM can also be used in the MLLM. The key additional technical difficulty in the MLLM is the evaluation of the data density. This can be accomplished by the following *extended GHK algorithm.*

Referring to the inequality constraints (6.2), let $\tilde{\boldsymbol{z}}_t = \boldsymbol{F}_t \tilde{\boldsymbol{y}}_t$. Without loss of generality, partition the q elements of $\boldsymbol{c}_t$, $\tilde{\boldsymbol{z}}_t$ and $\boldsymbol{d}_t$ each into q_{1t}, q_{2t} and q_{3t} components so that $\boldsymbol{c}_t' = (\boldsymbol{c}_{1t}', \boldsymbol{c}_{2t}', \boldsymbol{c}_{3t}')$, $\tilde{\boldsymbol{z}}_t' = (\tilde{\boldsymbol{z}}_{1t}', \tilde{\boldsymbol{z}}_{2t}', \tilde{\boldsymbol{z}}_{3t}')$ and $\boldsymbol{d}_t' = (\boldsymbol{d}_{1t}', \boldsymbol{d}_{2t}', \boldsymbol{d}_{3t}')$, with

$$\boldsymbol{c}_{1t} = \tilde{\boldsymbol{z}}_{1t} = \boldsymbol{d}_{1t}, \quad \boldsymbol{c}_{2t} \leqslant \tilde{\boldsymbol{z}}_{2t} \leqslant \boldsymbol{d}_{2t}, \quad -\infty \leqslant \tilde{\boldsymbol{z}}_{3t} \leqslant +\infty.$$

Then the data density for the set-valued outcome $\boldsymbol{z}_t = \{\tilde{\boldsymbol{z}}_t : \boldsymbol{c}_t \leqslant \tilde{\boldsymbol{z}}_t \leqslant \boldsymbol{d}_t\}$ may be expressed

$$p(\boldsymbol{z}_t|\theta) = p(\tilde{\boldsymbol{z}}_{1t}|\theta) \cdot P(\boldsymbol{c}_{2t} \leqslant \tilde{\boldsymbol{z}}_{2t} \leqslant \boldsymbol{d}_{2t}|\tilde{\boldsymbol{z}}_{1t}, \theta), \tag{B.13}$$

where θ is the vector of model parameters and any other relevant conditioning information. The construction of $\boldsymbol{z}_t$ and $\tilde{\boldsymbol{z}}_t$, and the partition of $\boldsymbol{c}_t$, $\tilde{\boldsymbol{z}}_t$ and $\boldsymbol{d}_t$ depends on the data, and not on the distributional assumptions or other features of the model.

Consequently evaluating the density of z_t, as opposed to $\boldsymbol{y}_t$, will change the data density by the same multiplicative factor in all models and thus leave the marginal likelihood unaffected. The partition of $\tilde{z}_t$ may be different from one observation to the next: in particular, $\tilde{z}_{3t}$ will arise when observations on one or more components of $\boldsymbol{y}_t$ are completely missing. Since the data density depends only on the first two components of $\tilde{z}_t$, it is convenient to denote $\tilde{z}'_{*t} = (\tilde{z}'_{1t}, \tilde{z}'_{2t})$.

With normally distributed disturbances, conditional on θ, $\tilde{z}_{*t} \sim N(\mu, \Sigma)$ with μ and Σ both known. Let $\Sigma = \boldsymbol{T}\boldsymbol{T}'$, where $\boldsymbol{T}$ is the unique lower triangular Choleski factor of Σ, and write

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \tilde{z}_{*t} - \mu = \begin{pmatrix} \tilde{z}_{1t} - \mu_1 \\ \tilde{z}_{2t} - \mu_2 \end{pmatrix} = \begin{bmatrix} \boldsymbol{T}_{11} & \boldsymbol{0} \\ \boldsymbol{T}_{21} & \boldsymbol{T}_{22} \end{bmatrix} \begin{pmatrix} \eta_{1t} \\ \eta_{2t} \end{pmatrix} = \boldsymbol{T}\eta_t. \tag{B.14}$$

Then $\tilde{z}_{1t} = \boldsymbol{c}_{1t}$ is equivalent to $\eta_{1t} = \boldsymbol{T}_{11}^{-1}(\boldsymbol{c}_{1t} - \mu_1)$ and

$$p(\tilde{z}_{1t} = \boldsymbol{c}_{1t}) = (2\pi)^{-q_1/2} \prod_{i=1}^{q_1} t_{ii}^{-1} \exp\left(-\eta'_{1t}\eta_{1t}/2\right).$$

Likewise $\boldsymbol{c}_{2t} \leqslant \tilde{z}_{2t} \leqslant \boldsymbol{d}_{2t}$, given $\tilde{z}_{1t} = \boldsymbol{c}_{1t}$, is equivalent to

$$\boldsymbol{c}_2 - \mu_2 - \boldsymbol{T}_{21}\eta_{1t} \leqslant \boldsymbol{T}_{22}\eta_{2t} \leqslant \boldsymbol{d}_2 - \mu_2 - \boldsymbol{T}_{21}\eta_{1t}. \tag{B.15}$$

The random vector in Equation (B.15) is η_{2t}, and the probability of the event (B.15) is the second component on the right side of Equation (B.13). This probability may be approximated by the GHK algorithm discussed in Section 2.1.

In the case of the multivariate Student-t distribution, $\tilde{z}_{*t} \sim t(\mu, \Sigma;\ \lambda)$. The relations (B.14) still obtain, except that now $\eta_t \sim T(\boldsymbol{0}, \boldsymbol{I}_{q_1+q_2};\ \lambda)$. The marginal distribution of $\tilde{z}_{1t}$ is $\tilde{z}_{1t} \sim t(\mu_1, \boldsymbol{T}_{11}\boldsymbol{T}'_{11};\ \lambda)$ so that

$$p(\tilde{z}_{1t} = \boldsymbol{c}_{1t}) = \Gamma\left[(\lambda + q_1)/2\right] \Gamma\left(\frac{\lambda}{2}\right)^{-1} (\lambda\pi)^{q_1/2} \prod_{i=1}^{q_1} t_{ii}^{-1} \left(1 + \eta'_{1t}\eta_{1t}/\lambda\right)^{-(\lambda+q_1)/2}.$$

The conditional distribution of $\tilde{z}_{2t}$ is also Student-t, with location vector

$$\mu_2 - \Sigma_{21}\Sigma_{11}^{-1}(\tilde{z}_{1t} - \mu_1) = \boldsymbol{T}_{21}\boldsymbol{T}_{11}^{-1}(\boldsymbol{c}_{1t} - \mu_1) = \mu_2 + \boldsymbol{T}_{21}\eta_{1t}.$$

The scale matrix is

$$\begin{aligned} &(1 + q_1/\lambda)^{-1}\left[1 + (\boldsymbol{c}_{1t} - \mu_1)'\Sigma_{11}^{-1}(\boldsymbol{c}_{1t} - \mu_1)/\lambda\right]\left(\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right) \\ &\quad = \left(\frac{\lambda + \eta'_{1t}\eta_{1t}}{\lambda + q_1}\right)\boldsymbol{T}_{22}\boldsymbol{T}'_{22}, \end{aligned}$$

and the degrees of freedom parameter is $\lambda + q_1$[39]. Hence the probability we seek is that of the event

$$c_2 - \mu_2 - T_{21}\eta_{1t} \leqslant \left(\frac{\lambda + \eta_{1t}'\eta_{1t}}{\lambda + q_1}\right)^{1/2} T_{22}\eta_{2t} \leqslant d_2 - \mu_2 - T_{21}\eta_{1t} \tag{B.16}$$

where $\eta_{2t} \sim t(\mathbf{0}, \boldsymbol{I}_{q_2};\ \lambda + q_1)$. By construction, $\eta_{2t} = w^{-1/2}\eta_{2t}^*$, where w and η_{2t}^* are independent, $(\lambda + q_1)\, w \sim \chi^2(\lambda + q_1)$, and $\eta_{2t}^* \sim N(\mathbf{0}, \boldsymbol{I}_{q_2})$. Hence the probability of the event (B.16) is the same as that of the event

$$c_2 - \mu_2 - T_{21}\eta_{1t} \leqslant T_{22}\eta_{2t}^{**} \leqslant d_2 - \mu_2 - T_{21}\eta_{1t}, \tag{B.17}$$

where first $w^* \sim \chi^2(\lambda + q_1)$ and then

$$\eta_{2t}^{**} \sim N\left\{\mathbf{0}, \left[\left(\lambda + \eta_{1t}'\eta_{1t}\right)/w^*\right] \cdot \boldsymbol{I}_{q_2}\right\}. \tag{B.18}$$

The probability of Equation (B.17) with η_{2t}^{**} having distribution (B.18), can be approximated by the GHK algorithm.

In the mixed normal distribution, $\tilde{z}_t \sim N(\mu^j, \sigma_j^2 \boldsymbol{I}_{q_1+q_2})$ with probability p_j. Let $\eta_{t1}^j = T_{11}^{-1}(\tilde{c}_{1t} - \mu_1^j)$. Using this notation and that developed previously,

$$p\left(\tilde{z}_{1t} = c_{1t}\right) = (2\pi)^{-q_1/2} \prod_{i=1}^{q_1} t_{ii}^{-1} \sum_{j=1}^{m} p_j \sigma_j^{-q_1} \exp\left(-\eta_{1t}^{j\prime}\eta_{1t}^{j}/2\sigma_j^2\right).$$

Conditional on $\tilde{z}_{1t} = c_{1t}$ and state j, $P_j(c_{2t} \leqslant \tilde{z}_{2t} \leqslant d_{2t})$ is the probability of the event

$$c_{2t} - \mu_2^j - T_{21}\eta_{t1}^j \leqslant T_{22}\eta_{2t}^j \leqslant d_{2t} - \mu_2^j - T_{21}\eta_{t1}^j,$$

where $\eta_{2t}^j \sim N(\mathbf{0}, \sigma_j^2 \boldsymbol{I}_{q_2})$. This probability can be approximated using the GHK algorithm. Removing the conditioning on state j,

$$p\left(z_{1t}\right) = p\left(\tilde{z}_{1t} = c_{1t}\right) \cdot \sum_{j=1}^{m} \tilde{p}_j P_j\left(c_{2t} \leqslant \tilde{z}_{2t} \leqslant d_{2t}\right),$$

where $\tilde{p}_j \propto p_j \sigma_j^{-1} \exp\left(-\eta_{t1}^{j\prime}\eta_{t1}^{j}/2\sigma_j^2\right)$ and $\sum_{j=1}^{m} \tilde{p}_j = 1$.[40]

[39] On conditional distributions in the multivariate Student-t distribution, see Johnson and Kotz (1972, pp. 134–135).

[40] An alternative algorithm is simply to apply the algorithm for the normal distribution to each state, then weight the outcomes by the probability parameters p_j. The method described in the text is more efficient – sometimes by a factor of ten or more – when combined with optimal allocation of simulation over states to minimize the time required to achieve a specified standard error of approximation of $\log\left[p(z_t)\right]$.

References

Albright, R., S. Lerman and C.F. Manski (1977), “Report on the development of an estimation program for the multinomial probit model”, Report prepared by Cambridge Systematics for the Federal Highway Administration.

Amemiya, T. (1985), Advanced Econometrics (Harvard University Press, Cambridge).

Andrews, D.W.K. (1999), “An improved simulator for multivariate normal rectangle probabilities and their derivatives”, Working paper (Yale University).

Bellman, R. (1957), Dynamic Programming (Princeton University Press, Princeton).

Bellman, R., R. Kalaba and B. Kotkin (1963), “Polynomial approximation – a new computational technique in dynamic programming: allocation processes”, Mathematics of Computation 1:155–161.

Borsch-Supan, A., and V. Hajivassiliou (1993), “Smooth unbiased multivariate probability simulators for maximum likelihood estimation of limited dependent variable models”, Journal of Econometrics 58:347–368.

Bunch, D. (1991), “Estimability in the multinomial probit model”, Transportation Research B 25:1–12.

Bunke, O., and X. Milhaud (1998), “Asymptotic behavior of Bayes estimates under possibly incorrect models”, Annals of Statistics 26(2):617–644.

Chib, S. (1995), “Marginal likelihood from the Gibbs output”, Journal of the American Statistical Association 90:1313–1321.

Chib, S., and E. Greenberg (1995a), “Understanding the Metropolis\erndash;Hastings algorithm”, The American Statistician 49:327–335.

Chib, S., and E. Greenberg (1995b), “Hierarchical analysis of sur models with extensions to correlated serial errors and time-varying parameter models”, Journal of Econometrics 68:339–360.

Chib, S., and E. Greenberg (1996), “Markov chain Monte Carlo simulation methods in econometrics”, Econometric Theory 12:409–431.

Cogburn, R. (1972), “The central limit theorem for Markov processes”, in: Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability 2 (University of California Press, Berkeley) 485–512.

Cosslett, S.R. (1983), “Distribution-free maximum likelihood estimator of the binary choice model”, Econometrica 51:765–782.

Danise, B. (1985), “Parameter estimability in the multinomial probit model”, Transportation Research B 19:526–528.

Dennis Jr, J.E., and R.B. Schnabel (1983), Numerical Methods for Unconstrained Optimization and Nonlinear Equations (Prentice-Hall, Englewood Cliffs).

Elrod, T., and M. Keane (1995), “A factor-analytic probit model for representing the market structure in panel data”, Journal of Marketing Research 32:1–16.

Erdem, T., and M. Keane (1996), “Decision making under uncertainty: capturing dynamic brand choice processes in turbulent consumer goods markets”, Marketing Science 15(1):1–20.

Ferguson, T.S. (1983), “Bayesian density estimation by mixtures of normal distributions”, in: H. Rivizi and J. Rustagi, eds., Recent Advances in Statistics (Academic Press, New York) 287–302.

Gallant, A.R., and D.W. Nychka (1987), “Semi-nonparametric maximum likelihood estimation”, Econometrica 55:363–390.

Gelfand, A.E., and D.K. Dey (1994), “Bayesian model choice: asymptotics and exact calculations”, Journal of the Royal Statistical Society Series B 56:501–514.

Gelfand, A.E., and A.F.M. Smith (1990), “Sampling based approaches to calculating marginal densities”, Journal of the American Statistical Association 85:398–409.

Gelman, A. (1996), “Inference and monitoring convergence”, in: W.R. Gilks, S. Richardson and D.J. Spiegelhalter, eds., Markov Chain Monte Carlo in Practice (Chapman and Hall) 131–140.

Gelman, A., and D.B. Rubin (1992), “Inference from iterative simulation using multiple sequences”, Statistical Science 7:457–472.

Gelman, A., J.B. Carlin, H.S. Stern and D.B. Rubin (1995), Bayesian Data Analysis (Chapman and Hall, London).

Geman, S., and D. Geman (1984), "Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images", IEEE Transactions on Pattern Analysis and Machine Intelligence 6:721–741.

Geweke, J. (1988), "Antithetic acceleration of Monte Carlo integration in Bayesian inference", Journal of Econometrics 38:73–90.

Geweke, J. (1989), "Bayesian inference in econometric models using Monte Carlo integration", Econometrica 57:1317–1340.

Geweke, J. (1991), "Efficient simulation from the multivariate normal and student-t distributions subject to linear constraints", in: E. M. Keramidas, ed., Computing Science and Statistics: Proceedings of the Twenty-Third Symposium on the Interface (Interface Foundation of North America, Inc., Fairfax) 571–578.

Geweke, J. (1992), "Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments", in: J.O. Berger, J.M. Bernardo, A.P. Dawid and A.F.M. Smith, eds., Proceedings of the Fourth Valencia International Meeting on Bayesian Statistics (Oxford University Press, Oxford) 169–194.

Geweke, J. (1993), "Bayesian treatment of the independent Student-t linear model", Journal of Applied Econometrics 8:S19–S40.

Geweke, J. (1996), "Monte Carlo simulation and numerical integration", in: H.M. Amman, D.A. Kendrick and J. Rust, eds., Handbook of Computational Economics (Amsterdam, North-Holland) 731–800.

Geweke, J. (1999), "Using simulation methods for Bayesian econometric models: inference, development, and communication (with discussion and reply)", Econometric Reviews 18:1–127.

Geweke, J., and M. Keane (1995), "Bayesian inference for dynamic discrete choice models without the need for dynamic programming", Working paper (Federal Reserve Bank of Minneapolis). Also in: Mariano, Schuermann and Weeks, eds., Simulation Based Inference and Econometrics: Methods and Applications (Cambridge University Press, Cambridge) forthcoming.

Geweke, J., and M. Keane (1999), "Mixture of normals probit models", in: C. Hsiao, K. Lahiri, L.-F. Lee and H. Pesaran, eds., Analysis of Panels and Limited Dependent Variable Models: An Edited Volume in Honor of G.S. Maddala (Cambridge University Press) 49–78.

Geweke, J., and M. Keane (2000), "An empirical analysis of earnings dynamics among men in the PSID: 1968–1989", Journal of Econometrics 92:293–356.

Geweke, J., M. Keane and D.E. Runkle (1994), "Alternative computational approaches to statistical inference in the multinomial probit model", Review of Economics and Statistics 76(4):609–632.

Geweke, J., M. Keane and D.E. Runkle (1997), "Statistical inference in the multinomial multiperiod probit model", Journal of Econometrics 80:125–165.

Geweke, J., D. Houser and M. Keane (1998), "Simulation based inference for dynamic multinomial choice models", in: B.H. Baltaji, ed., Companion for Theoretical Econometrics (Basil Blackwell, London) forthcoming.

Geweke, J., W. McCausland and J. Stevens (2000), "Using simulation methods for Bayesian econometric models", in: D. Giles, ed., Computer Aided Econometrics (Marcel Dekker, New York) forthcoming.

Geyer, C.J. (1992), "Practical Markov chain Monte Carlo", Statistical Science 7:473–481.

Goldberger, A.S. (1991), A Course in Econometrics (Cambridge, Harvard University Press).

Greene, W.H. (1997), Econometric Analysis, 3rd edition (Prentice Hall, Upper Saddle River).

Hajivassiliou, V. (1991), "Simulation estimation methods for limited dependent variable models," Cowles Foundation discussion paper 1007 (Cowles Foundation for Research in Economics, Yale University).

Hajivassiliou, V., and P.A. Ruud (1994), "Classical estimation methods for ldv models using simulation", in R.F. Engle and D.L. McFadden eds., Handbook of Econometrics, vol IV (Amsterdam, Elsevier) 2384-2443.

Hajivassiliou, V., D. McFadden and P.A. Ruud (1996), "Simulation of multivariate normal rectangle probabilities and their derivatives: theoretical and computational results", Journal of Econometrics 72:85–134.

Hammersly, J.M., and D.C. Handscomb (1964), Monte Carlo Methods (Methuen, London).

Hannan, E.J. (1970), Multiple Time Series Analysis (Wiley, New York).

Hastings, W.K. (1970), "Monte Carlo sampling methods using Markov chains and their applications", Biometrika 57:97–109.

Heckman, J.J. (1979), "Sample selection bias as a specification error", Econometrica 47:153–161.

Heckman, J.J. (1981), "Statistical models for discrete panel data", in: C. Manski and D. McFadden, eds., Structural Analysis of Discrete Data with Econometric Applications (MIT Press, Cambridge).

Heckman, J.J., and G. Sedlacek (1985), "Heterogeneity, aggregation and market wage functions: an empirical model of self-selection in the labor market", Journal of Political Economy 93:1077–1125.

Horowitz, J.L. (1992), "A smoothed maximum score estimator for the binary response model", Econometrica 60:505–531.

Hotz, V.J., and R.A. Miller (1993), "Conditional choice probabilities and the estimation of dynamic programming models", Review of Economic Studies 60:497–530.

Houser, D. (1998), "Bayesian analysis of a dynamic, stochastic model of labor supply and saving", Working paper (University of Arizona).

Huber, P. (1967), "The behavior of maximum likelihood estimates under nonstandard conditions", in: Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability 1 (University of California Press, Berkeley) 221–233.

Ichimura, H. (1993), "Semiparametric least squares (SLS) and weighted SLS estimation of single-index models", Journal of Econometrics 58:71–120.

IMSL (1994), IMSL Math Library: FORTRAN Subroutines for Mathematical Applications (Visual Numerics, Inc., Houston).

Johnson, N.L., and S. Kotz (1972), Distributions in Statistics: Continuous Multivariate Distributions (Wiley, New York).

Johnson, N.L., S. Kotz and N. Balakrishnan (1995), Continuous Univariate Distributions, Vol. 2, 2nd edition (Wiley, New York).

Judd, K.L. (1992), "Projection methods for solving aggregate growth models", Journal of Economic Theory 58:410–452.

Keane, M. (1990), "Four essays in empirical macro and labor economics", Ph.D. dissertation (Brown University).

Keane, M. (1992), "A note on identification in the multinomial probit model", Journal of Business and Economic Statistics 10(2):192–200.

Keane, M. (1993), "Simulation estimation for panel data models with limited dependent variables", in: G.S. Maddala, C.R. Rao and H.D. Vinod, eds., The Handbook of Statistics (North-Holland, Amsterdam) 545–572.

Keane, M. (1994), "A computationally practical simulation estimator for panel data", Econometrica 62(1):95–116.

Keane, M. (1997), "Modeling heterogeneity and state dependence in consumer choice behavior", Journal of Business and Economic Statistics 15(3):310–327.

Keane, M., and D.E. Runkle (1990), "Testing the rationality of price forecasts: new evidence from panel data", American Economic Review 80(4):714–735.

Keane, M., and K.I. Wolpin (1994), "The solution and estimation of discrete choice dynamic programming models by simulation: Monte Carlo evidence", Review of Economics and Statistics 76(4):648–672.

Keane, M., and K.I. Wolpin (1997), "The career decisions of young men", Journal of Political Economy 105(3):473–522.

Keane, M., and K.I. Wolpin (2000a), "Equalizing race differences in school attainment and labor market success", Journal of Labor Economics 18:614–652.

Keane, M., and K.I. Wolpin (2000b), "The effect of parental transfers and borrowing constraints on educational attainment", International Economic Review, forthcoming.

Kiefer, J., and J. Wolfowitz (1956), "Consistency of the maximumm likelihood estimator in the presence of infinitely many incidental parameters", Annals of Mathematical Statistics 27:887–906.

Klein, R.W., and R.H. Spady (1993), “An efficient semiparametric estimator for binary response models”, Econometrica 61:387–421.

Kloek, T., and H.K. van Dijk (1978), “Bayesian estimates of equation system parameters: an application of integration by Monte Carlo”, Econometrica 46:1–20.

Lancaster, T. (1997), “Exact structural inference in optimal job search models”, Journal of Business and Economic Statistics 15(2):165–179.

Lee, L.-F. (1978), “Unionism and wage rates: a simultaneous equation model with qualitative and limited dependent variables”, International Economic Review 19:415–433.

Lee, L.-F. (1979), “Identification and estimation in binary choice models with limited (censored) dependent variables”, Econometrica 47:977–996.

Lee, L.-F. (1992), “On efficiency of methods of simulated moments and maximum simulated likelihood estimation of discrete response models”, Econometric Theory 8:518–552.

Lee, L.-F. (1995), “Asymptotic bias in maximum simulated likelihood estimation of discrete choice models”, Econometric Theory 11:437–483.

Lee, L.-F. (1997), “Simulated maximum likelihood estimation of dynamic discrete choice statistical models: some Monte Carlo results”, Journal of Econometrics 82:1–35.

Lerman, S., and C.F. Manski (1981), “On the use of simulated frequencies to approximate choice probabilities”, in: C.F. Manski and D. McFadden, eds., Structural Analysis of Discrete Data with Econometric Applications (MIT Press, Cambridge).

Lewbel, A. (1997), “Semiparametric estimation of location and other discrete choice moments”, Econometric Theory 13:32–51.

Lillard, L.A., and R.J. Willis (1978), “Dynamic aspects of earnings mobility”, Econometrica 46: 985–1012.

Maddala, G.S. (1992), Introduction to Econometrics, 2nd edition (Macmillan, New York).

Manski, C.F. (1985), “Semiparametric analysis of discrete response: asymptotic properties of the maximum score estimator”, Journal of Econometrics 27:313–333.

Manski, C.F. (1991), “Nonparametric estimation of expectations in the analysis of discrete choice under uncertainty”, in: W. Barnett, J. Powell and G. Tauchen, eds., Nonparametric and Semiparametric Methods in Econometrics and Statistics (Cambridge University Press, Cambridge).

Marcet, A. (1994), “Simulation analysis of dynamic stochastic models: application to theory and estimation”, in: C. Sims, ed., Advances in Econometrics, Sixth World Congress, Vol. II (Cambridge University Press, Cambridge) 81–118.

McFadden, D. (1989), “A method of simulated moments for estimation of multinomial probits without numerical integration”, Econometrica 57:995–1026.

Mengersen, K.L., and R.L. Tweedie (1996), “Rates of convergence of the Hastings and Metropolis algorithms,” Annals of Statistics 24:101–121.

Metropolis, N., A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller and E. Teller (1953), “Equation of state calculations by fast computing machines”, The Journal of Chemical Physics 21:1087–1092.

Pakes, A. (1986), “Patents as options: some estimates of the value of holding European patent options,” Econometrica 54:755–785.

Pakes, A., and D. Pollard (1989), “Simulation and the asymptotics of optimization estimators”, Econometrica 57(5):1027–1058.

Peskun, P.H. (1973), “Optimum Monte-Carlo sampling using Markov chains”, Biometrika 60:607–612.

Pfanzagl, J. (1969), “On the measurability and consistency of minimum contrast estimators”, Metrika 14:249–272.

Powell, J.L., J.H. Stock and T.M. Stoker (1989), “Semiparametric estimation of index coefficients”, Econometrica 57:1403–1430.

Ripley, R.D. (1987), Stochastic Simulation (Wiley, New York).

Roberts, G.O., and A.F.M. Smith (1994), “Simple conditions for the convergence of the Gibbs sampler and Metropolis–Hastings algorithms”, Stochastic Processes and Their Applications 49:207–216.

Roy, A.D. (1951), “Some thoughts on the distribution of earnings”, Oxford Economics Papers 3:135–146.

Rust, J. (1987), "Optimal replacement of GMC bus engines: an empirical model of Harold Zurcher", Econometrica 55:999–1033.

Rust, J. (1997), "Using randomization to break the curse of dimensionality", Econometrica 65:487–516.

Santos, M.S., and J. Vigo-Aguiar (1998), "Analysis of a numerical dynamic programming algorithm applied to economic models", Econometrica 66:409–426.

Stern, S. (1991), "Approximate solutions to stochastic dynamic programming problems", Mimeo (University of Virginia).

Tanner, M.A., and W.H. Wong (1987), "The calculation of posterior distributions by data augmentation", Journal of the American Statistical Association 82:528–550.

Tapia, R.A., and J.R. Thompson (1978), Nonparametric Probability Density Estimation (Johns Hopkins University Press, Baltimore).

Tierney, L. (1994), "Markov chains for exploring posterior distributions (with discussion and rejoinder)", Annals of Statistics 22:1701–1762.

Traub, J.F., G.W. Wasilkowski and H. Wozniakowski (1988), Information-Based Complexity (Academic Press, Amsterdam).

Vijverberg, W.P.M. (1997), "Monte Carlo evaluation of multivariate normal probabilities", Journal of Econometrics 76:281–307.

Zeger, S.L., and M.R. Karim (1991), "Generalized linear models with random effects: a Gibbs sampling approach", Journal of the American Statistical Association 86:79–86.

Zellner, A. (1962), "An efficient method of estimating seemingly unrelated regressions and tests of aggregation bias", Journal of the American Statistical Association 57:500–509.

Zellner, A. (1971), An Introduction to Bayesian Inference in Econometrics (Wiley, New York).

Zellner, A., and C. Min (1995), "Gibbs sampler convergence criteria", Journal of the American Statistical Association 90:921–927.

Chapter 57

MARKOV CHAIN MONTE CARLO METHODS: COMPUTATION AND INFERENCE

SIDDHARTHA CHIB *

John M. Olin School of Business, Washington University, Campus Box 1133, 1 Brookings Dr.,
St. Louis, MO 63130, USA

Contents

Abstract 3570
Keywords 3570
1. Introduction 3571
 1.1. Organization 3573
2. Classical sampling methods 3573
 2.1. Inverse transform method 3573
 2.2. Accept–reject algorithm 3575
 2.3. Method of composition 3576
3. Markov chains 3576
 3.1. Definitions and results 3577
 3.2. Computation of numerical accuracy and inefficiency factor 3579
4. Metropolis–Hastings algorithm 3580
 4.1. The algorithm 3581
 4.2. Convergence results 3584
 4.3. Example 3585
 4.4. Multiple-block M–H algorithm 3587
5. The Gibbs sampling algorithm 3589
 5.1. The algorithm 3590
 5.2. Connection with the multiple-block M–H algorithm 3591
 5.3. Invariance of the Gibbs Markov chain 3592
 5.4. Sufficient conditions for convergence 3592
 5.5. Estimation of density ordinates 3592
 5.6. Example: simulating a truncated multivariate normal 3594
6. Sampler performance and diagnostics 3595
7. Strategies for improving mixing 3596
 7.1. Choice of blocking 3597

* email: chib@olin.wustl.edu

Handbook of Econometrics, Volume 5, Edited by J.J. Heckman and E. Leamer

7.2. Tuning the proposal density 3597
7.3. Other strategies 3598
8. MCMC algorithms in Bayesian estimation 3599
8.1. Overview 3599
8.2. Notation and assumptions 3600
8.3. Normal and student-*t* regression models 3602
8.4. Binary and ordinal probit 3604
8.5. Tobit censored regression 3607
8.6. Regression with change point 3608
8.7. Autoregressive time series 3610
8.8. Hidden Markov models 3612
8.9. State space models 3614
8.10. Stochastic volatility model 3616
8.11. Gaussian panel data models 3619
8.12. Multivariate binary data models 3620
9. Sampling the predictive density 3623
10. MCMC methods in model choice problems 3626
10.1. Background 3626
10.2. Marginal likelihood computation 3627
10.3. Model space-parameter space MCMC algorithms 3633
10.4. Variable selection 3636
10.5. Remark 3639
11. MCMC methods in optimization problems 3639
12. Concluding remarks 3641
References 3642

Abstract

This chapter reviews the recent developments in Markov chain Monte Carlo simulation methods. These methods, which are concerned with the simulation of high dimensional probability distributions, have gained enormous prominence and revolutionized Bayesian statistics. The chapter provides background on the relevant Markov chain theory and provides detailed information on the theory and practice of Markov chain sampling based on the Metropolis–Hastings and Gibbs sampling algorithms. Convergence diagnostics and strategies for implementation are also discussed. A number of examples drawn from Bayesian statistics are used to illustrate the ideas. The chapter also covers in detail the application of MCMC methods to the problems of prediction and model choice.

Keywords

JEL classification: C1, C4

1. Introduction

This chapter is concerned with the theory and practice of Markov chain Monte Carlo (MCMC) simulation methods. These methods which deal with the simulation of high dimensional probability distributions, have over the last decade gained enormous prominence, sparked intense research interest, and energized Bayesian statistics [Tanner and Wong (1987), Casella and George (1992), Gelfand and Smith (1990, 1992), Smith and Roberts (1993), Tierney (1994), Chib and Greenberg (1995a, 1996), Besag, Green, Higdon and Mengersen (1995), Albert and Chib (1996), Tanner (1996), Gilks, Richardson and Spiegelhalter (1996), Carlin and Louis (2000), Geweke (1997), Gammerman (1997), Brooks (1998), Robert and Casella (1999)]. The idea behind these methods is simple and extremely general. In order to sample a given probability distribution that is referred to as the target distribution, a suitable Markov chain is constructed with the property that its limiting, invariant distribution is the target distribution. Depending on the specifics of the problem, the Markov chain can be constructed by the Metropolis–Hastings algorithm, the Gibbs sampling method, a special case of the Metropolis method, or hybrid mixtures of these two algorithms. Once the Markov chain has been constructed, a sample of (correlated) draws from the target distribution can be obtained by simulating the Markov chain a large number of times and recording its values. In many situations, Markov chain Monte Carlo simulation provides the only practical way of obtaining samples from high dimensional probability distributions.

Markov chain sampling methods originated with the work of Metropolis, Rosenbluth, Rosenbluth, Teller and Teller (1953) who proposed an algorithm to simulate a high dimensional discrete distribution. This algorithm found wide application in statistical physics but was mostly unknown to statisticians until the paper of Hastings (1970). Hastings generalized the Metropolis algorithm and applied it to the simulation of discrete and continuous probability distributions such as the normal and Poisson. Outside of statistical physics, Markov chain methods first found applications in spatial statistics and image analysis [Besag (1974)]. The more recent interest in MCMC methods can be traced to the papers of Geman and Geman (1984), who developed an algorithm that later came to be called the Gibbs sampler, to sample a discrete distribution, Tanner and Wong (1987), who proposed a MCMC scheme involving "data augmentation" to sample posterior distributions in missing data problems, and Gelfand and Smith (1990), where the value of the Gibbs sampler was demonstrated for general Bayesian inference with continuous parameter spaces.

In Bayesian applications, the target distribution is typically the posterior distribution of the parameters, given the data. If $\mathcal{M}$ denotes a particular model, $p(\boldsymbol{\psi}|\mathcal{M})$, $\boldsymbol{\psi} \in \Re^d$, the prior density of the parameters in that model and $f(\boldsymbol{y}|\boldsymbol{\psi}, \mathcal{M})$ the assumed sampling density (likelihood function) for a vector of observations $\boldsymbol{y}$, then the posterior density is given by

$$\pi(\boldsymbol{\psi}|\boldsymbol{y}, \mathcal{M}) \propto p(\boldsymbol{\psi}|\mathcal{M}) f(\boldsymbol{y}|\boldsymbol{\psi}, \mathcal{M}), \tag{1}$$

where the normalizing constant of the density, called the marginal likelihood,

$$m(\boldsymbol{y}|\mathcal{M}) = \int_{\Re^d} p(\boldsymbol{\psi}|\mathcal{M}) f(\boldsymbol{y}|\boldsymbol{\psi}, \mathcal{M})\,\mathrm{d}\boldsymbol{\psi},$$

is almost never known in analytic form. As may be expected, an important goal of the Bayesian analysis is to summarize the posterior density. Particular summaries, such as the posterior mean and posterior covariance matrix, are especially important as are interval estimates (called credible intervals) with specified posterior probabilities. The calculation of these quantities reduces to the evaluation of the following integral

$$\int_{\Re^d} h(\boldsymbol{\psi})\,\pi(\boldsymbol{\psi}|\boldsymbol{y}, \mathcal{M})\,\mathrm{d}\boldsymbol{\psi},$$

under various choices of the function h. For example, to get the posterior mean, one lets $h(\boldsymbol{\psi}) = \boldsymbol{\psi}$ and for the second moment matrix one lets $h(\boldsymbol{\psi}) = \boldsymbol{\psi}\boldsymbol{\psi}'$, from which the posterior covariance matrix and posterior standard deviations may be computed.

In the pre MCMC era, posterior summaries were usually obtained either by analytic approximations, such as the method of Laplace for integrals [Tierney and Kadane (1986)], or by the method of importance sampling [Kloek and van Dijk (1978), Geweke (1989)]. Although both techniques continue to have uses (for example, the former in theoretical, asymptotic calculations), neither method is sufficiently flexible to be used routinely for the kinds of high-dimensional problems that arise in practice. A shift in thinking was made possible by the advent of MCMC methods. Instead of focusing on the question of moment calculation directly one may consider the more general question of drawing sample variates from the distribution whose summaries are sought. For example, to summarize the posterior density $\pi(\boldsymbol{\psi}|\boldsymbol{y}, \mathcal{M})$ one can produce a simulated sample $\{\boldsymbol{\psi}^{(1)}, \ldots, \boldsymbol{\psi}^{(M)}\}$ from this posterior density, and from this simulated sample, the posterior expectation of $h(\boldsymbol{\psi})$ can be estimated by the average

$$M^{-1} \sum_{j=1}^{M} h(\boldsymbol{\psi}^{(j)}). \tag{2}$$

Under independent sampling from the posterior, which is rarely feasible, this calculation would be justified by classical laws of large numbers. In the context of MCMC sampling the draws are correlated but, nonetheless, a suitable law of large numbers for Markov chains that is presented below can be used establish the fact that

$$M^{-1} \sum_{j=1}^{M} h(\boldsymbol{\psi}^{(j)}) \to \int_{\Re^d} h(\boldsymbol{\psi})\,\pi(\boldsymbol{\psi}|\boldsymbol{y}, \mathcal{M})\,\mathrm{d}\boldsymbol{\psi}, \quad M \to \infty.$$

It is important to bear in mind that the convergence specified here is in terms of the simulation sample size M and not in terms of the data sample size n which is fixed.

This means that one can achieve any desired precision by taking M to be as large as required, subject to the constraint on computing time.

The Monte Carlo approach to inference also provides elegant solutions to the Bayesian problems of prediction and model choice. For the latter, algorithms are available that proceed to sample over both model space and parameter space, such as in the methods of Carlin and Chib (1995) and Green (1995), or those that directly compute the evidential quantities that are required for Bayesian model comparisons, namely marginal likelihoods and their ratios, Bayes factors [Jeffreys (1961)]; these approaches are developed by Gelfand and Dey (1994), Chib (1995), Verdinelli and Wasserman (1995), Meng and Wong (1996), DiCiccio, Kass, Raftery and Wasserman (1997), Chib and Jeliazkov (2001), amongst others. Discussion of these techniques is provided in detail below.

1.1. Organization

The rest of the chapter is organized as follows. Section 2 provides a brief review of three classical sampling methods that are discussed or used in the sequel. Section 3 summarizes the relevant Markov chain theory that justifies simulation by MCMC methods. In particular, we provide the conditions under which discrete-time and continuous state space Markov chains satisfy a law of large numbers and a central limit theorem. The Metropolis–Hastings algorithm is discussed in Section 4 followed by the Gibbs sampling algorithm in Section 5. Methods for diagnosing convergence are considered in Section 6 and strategies for improving the mixing of the Markov chains in Section 7. In Section 8 we discuss how MCMC methods can be applied to simulate the posterior distributions that arise in various canonical statistical models. Bayesian prediction and model choice problems are presented in Sections 9 and 10, respectively, and the MCMC-based EM algorithm is considered in Section 11. Section 12 concludes with brief comments about new and emerging directions in MCMC methods.

2. Classical sampling methods

We now briefly review three sampling methods, that we refer to as classical methods, that deliver independent and identically distributed draws from the target density. Authoritative surveys of these and other such methods are provided by Devroye (1985), Ripley (1987) and Gentle (1998). Although these methods are technically outside the scope of this chapter, the separation is somewhat artificial because, in practice, all MCMC methods in one way or another make some use of classical simulation methods. The ones we have chosen to discuss here are those that are mentioned or used explicitly in the sequel.

2.1. Inverse transform method

This method is particularly useful in the context of discrete distribution functions and is based on taking the inverse transform of the cumulative distribution function (hence its name). Suppose we want to generate the value of a discrete random variable with mass function

$$\Pr(\psi = \psi_j) = p_j,\ j = 1, 2, \ldots, \sum_j p_j = 1,$$

and cumulative mass function

$$\Pr(\psi \leqslant \psi_j) \equiv F(\psi_j) = p_1 + p_2 + \cdots + p_j.$$

The function F is a right-continuous stair function that has jumps at the point ψ_j equal to p_j and is constant otherwise. It is not difficult to see that its inverse takes the form

$$F^{-1}(u) = \psi_j \quad \text{if} \quad p_1 + \cdots + p_{j-1} \leqslant u \leqslant p_1 + \cdots + p_j. \tag{3}$$

A random variate from this distribution is obtained by generating U uniform on $(0, 1)$ and computing $F^{-1}(U)$ where F^{-1} is the inverse function in Equation (3). This method samples ψ_j with probability p_j because

$$\begin{aligned}\Pr(F^{-1}(U) = \psi_j) &= \Pr(p_1 + \cdots + p_{j-1} \leqslant U \leqslant p_1 + \cdots + p_j) \\ &= p_j.\end{aligned}$$

An equivalent version is available for continuous random variables. An important application is to the sampling of a truncated normal distribution. Suppose, for example, that

$$\psi \sim \mathcal{TN}_{(a,b)}(\mu, \sigma^2),$$

a univariate truncated normal distribution truncated to the interval (a, b), with distribution function

$$F(t) = \begin{cases} 0 & \text{if } \psi < a \\ \frac{1}{p_2 - p_1}\left(\Phi(\frac{t-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})\right) & \text{if } a < \psi < b \\ 1 & \text{if } b < \psi \end{cases}, \tag{4}$$

where

$$p_1 = \Phi\left(\frac{a-\mu}{\sigma}\right); \quad p_2 = \Phi\left(\frac{b-\mu}{\sigma}\right).$$

To generate a sample variate from this distribution one must solve the equation $F(t) = U$, where U is uniform on $(0, 1)$. Algebra yields

$$t = \mu + \sigma\Phi^{-1}\left(p_1 + U(p_2 - p_1)\right). \tag{5}$$

Although the inverse distribution method is useful it is rather difficult to apply in the setting of multi-dimensional distributions.

2.2. Accept–reject algorithm

The accept–reject method is the basis for many of the well known univariate random number generators that are provided in software programs. This method is characterized by a source density $h(\psi)$ which is used to supply candidate values and a constant c, that is determined by analysis, such that for all ψ

$$\pi(\psi) \leqslant ch(\psi).$$

Note that the accept–reject method does not require knowledge of the normalizing constant of π because that constant can be absorbed in c. Then, in the accept–reject method, one draws a variate from h, accepting it with probability $\pi(\psi)/\{ch(\psi)\}$. If the particular proposal is rejected, a new one is drawn and the process continued until one is accepted. The accepted draws constitute an independent and identically distributed (i.i.d.) sample from π.

In algorithmic form, the accept–reject method can be described as follows.

Algorithm 1: Accept–reject

```
(1) Repeat for $j = 1, 2, \dots, M$.
    (a) Generate
            $\psi' \sim h(\psi); \quad U \sim \mathrm{Unif}(0, 1).$
    (b) Let $\psi^{(j)} = \psi'$ if
            $U \leqslant \dfrac{\pi(\psi')}{ch(\psi')},$
        otherwise go to step 1(a).
(2) Return the values $\{\psi^{(1)}, \psi^{(2)}, \dots, \psi^{(M)}\}$.
```

The idea behind this algorithm may be explained quite simply using Figure 1. Imagine drawing random bivariate points in the region bounded above by the function $ch(\psi)$ and below by the x-axis. A point in this region may be drawn by first drawing ψ' from $h(\psi)$, which fixes the x-coordinate of the point, and then drawing the y-coordinate of the point as $Uch(\psi')$. Now, if $Uch(\psi') \leqslant \pi(\psi')$, the point lies below π and is accepted; but the latter is simply the acceptance condition of the AR method, which completes the justification.

Below we shall discuss a Markov chain Monte Carlo version of the accept–reject method that can be used when the condition $\pi(\psi) \leqslant ch(\psi)$ does not hold for all values of ψ.

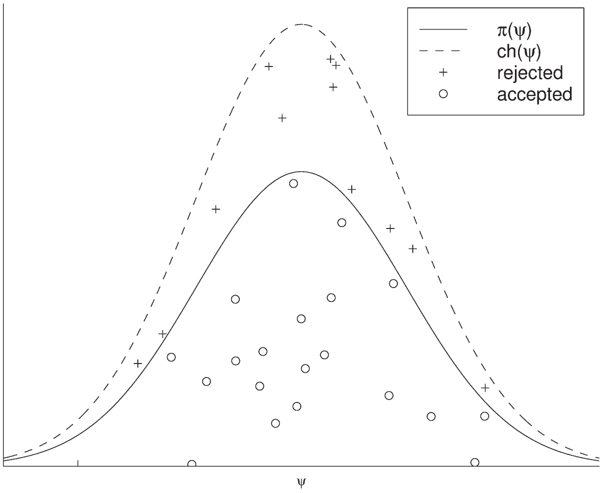

Fig. 1. Graphical illustration of the accept–reject method.

2.3. *Method of composition*

This method is based on the observation that if the joint density $\pi(\psi_1, \psi_2)$ is expressed as

$$\pi(\psi_1, \psi_2) = \pi(\psi_1)\, \pi(\psi_2 | \psi_1),$$

and each density on the right hand side is easily sampled, then a draw from the joint distribution may be obtained by
(1) drawing $\psi_1^{(j)}$ from $\pi(\psi_1)$ and then
(2) drawing $\psi_2^{(j)}$ from $\pi(\psi_2 | \psi_1^{(j)})$.

Because $(\psi_1^{(j)}, \psi_2^{(j)})$ is a draw from the joint distribution it follows that the second component of the simulated vector is a draw from the marginal distribution of ψ_2:

$$\psi_2^{(j)} \sim \pi(\psi_2) = \int \pi(\psi_2 | \psi_1)\, \pi(\psi_1)\, \mathrm{d}\psi_1.$$

Thus, to obtain a draw $\psi_2^{(j)}$ from $\pi(\psi_2)$, it is sufficient to produce a sample from the joint distribution and retain the second component. This method is quite important and arises frequently in the setting of MCMC methods.

3. Markov chains

Markov chain Monte Carlo is a method to sample a given multivariate distribution π^* by constructing a suitable Markov chain with the property that its limiting, invariant distribution, is the target distribution π^*. In most problems of interest, the distribution π^* is absolutely continuous and, as a result, the theory of MCMC methods is based on that of Markov chains on continuous state spaces outlined, for example, in Nummelin (1984) and Meyn and Tweedie (1993). Tierney (1994) is the fundamental reference for drawing the connections between this elaborate Markov chain theory and MCMC methods. Basically, the goal of the analysis is to specify conditions under which the constructed Markov chain converges to the invariant distribution, and conditions under which sample path averages based on the output of the Markov chain satisfy a law of large numbers and a central limit theorem.

3.1. Definitions and results

A Markov chain is a collection of random variables (or vectors) $\boldsymbol{\Phi} = \{\boldsymbol{\Phi}_i : i \in T\}$ where $T = \{0, 1, 2, \ldots\}$. The evolution of the Markov chain on a space $\Omega \subseteq \Re^p$ is governed by the *transition kernel*

$$\begin{aligned} P(\boldsymbol{x}, A) &\equiv \Pr(\boldsymbol{\Phi}_{i+1} \in A | \boldsymbol{\Phi}_i = \boldsymbol{x}, \boldsymbol{\Phi}_j, j < i) \\ &= \Pr(\boldsymbol{\Phi}_{i+1} \in A | \boldsymbol{\Phi}_i = \boldsymbol{x}), \quad \boldsymbol{x} \in \Omega, \quad A \subset \Omega, \end{aligned}$$

which embodies the Markov assumption that the distribution of each succeeding state in the sequence, given the current and the past states, depends only on the current state.

In general, in the context of Markov chain simulations, the transition kernel has both a continuous and a discrete component. For some function $p(\boldsymbol{x}, \boldsymbol{y}) : \Omega \times \Omega \to \Re^+$, the kernel can be expressed as

$$P(\boldsymbol{x}, \mathrm{d}\boldsymbol{y}) = p(\boldsymbol{x}, \boldsymbol{y})\, \mathrm{d}\boldsymbol{y} + r(\boldsymbol{x})\, \delta_{\boldsymbol{x}}(\mathrm{d}\boldsymbol{y}), \tag{6}$$

where $p(\boldsymbol{x}, \boldsymbol{x}) = 0$, $\delta_{\boldsymbol{x}}(\mathrm{d}\boldsymbol{y}) = 1$ if $\boldsymbol{x} \in \mathrm{d}\boldsymbol{y}$ and 0 otherwise, $r(\boldsymbol{x}) = 1 - \int_\Omega p(\boldsymbol{x}, \boldsymbol{y})\, \mathrm{d}\boldsymbol{y}$. This transition kernel specifies that transitions from $\boldsymbol{x}$ to $\boldsymbol{y}$ occur according to $p(\boldsymbol{x}, \boldsymbol{y})$ and transitions from $\boldsymbol{x}$ to $\boldsymbol{x}$ occur with probability $r(\boldsymbol{x})$.

The transition kernel is thus the distribution of $\boldsymbol{\Phi}_{i+1}$ given that $\boldsymbol{\Phi}_i = \boldsymbol{x}$. The nth-step-ahead transition kernel is given by

$$P^{(n)}(\boldsymbol{x}, A) = \int_\Omega P(\boldsymbol{x}, \mathrm{d}\boldsymbol{y})\, P^{(n-1)}(\boldsymbol{y}, A),$$

where $P^{(1)}(\boldsymbol{x}, \mathrm{d}\boldsymbol{y}) = P(\boldsymbol{x}, \mathrm{d}\boldsymbol{y})$ and

$$P(\boldsymbol{x}, A) = \int_A P(\boldsymbol{x}, \mathrm{d}\boldsymbol{y}). \tag{7}$$

The objective is to elucidate the conditions under which the nth iterate of the transition kernel converges to the invariant distribution π^* as $n \to \infty$. The invariant distribution satisfies

$$\pi^*(\mathrm{d}\boldsymbol{y}) = \int_\Omega P(\boldsymbol{x}, \mathrm{d}\boldsymbol{y})\, \pi(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}, \tag{8}$$

where π is the density of π^* with respect to the Lebesgue measure (thus, $\pi^*(\mathrm{d}\boldsymbol{y}) = \pi(\boldsymbol{y})\, \mathrm{d}\boldsymbol{y}$). The invariance condition states that if $\boldsymbol{\Phi}_i$ is distributed according to π^*, then all subsequent elements of the chain are also distributed as π^*. It should be noted that Markov chain samplers are invariant by construction and therefore the existence of the invariant distribution does not have to be checked in any particular application of MCMC methods.

A Markov chain is said to be *reversible* if the function $p(\boldsymbol{x},\boldsymbol{y})$ in Equation (6) satisfies

$$f(\boldsymbol{x})p(\boldsymbol{x},\boldsymbol{y}) = f(\boldsymbol{y})p(\boldsymbol{y},\boldsymbol{x}), \tag{9}$$

for a density $f(\cdot)$. If this condition holds, it can be shown that $f(\cdot) = \pi(\cdot)$. A reversible chain has π^* as an invariant distribution [see Tierney (1994)]. To verify this we evaluate the right hand side of Equation (8):

$$\begin{aligned}\int P(\boldsymbol{x},A)\,\pi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} &= \int\left\{\int_A p(\boldsymbol{x},\boldsymbol{y})\,\mathrm{d}\boldsymbol{y}\right\}\pi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} + \int r(\boldsymbol{x})\,\delta_x(A)\,\pi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x},\\ &= \int_A\left\{\int p(\boldsymbol{x},\boldsymbol{y})\,\pi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\right\}\mathrm{d}\boldsymbol{y} + \int_A r(\boldsymbol{x})\,\pi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x},\\ &= \int_A\left\{\int p(\boldsymbol{y},\boldsymbol{x})\,\pi(\boldsymbol{y})\,\mathrm{d}\boldsymbol{x}\right\}\mathrm{d}\boldsymbol{y} + \int_A r(\boldsymbol{x})\,\pi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x},\\ &= \int_A (1 - r(\boldsymbol{y}))\,\pi(\boldsymbol{y})\,\mathrm{d}\boldsymbol{y} + \int_A r(\boldsymbol{x})\,\pi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x},\\ &= \int_A \pi(\boldsymbol{y})\,\mathrm{d}\boldsymbol{y}.\end{aligned} \tag{10}$$

A minimal requirement to ensure that the Markov chain satisfies a law of large numbers is that of π^**-irreducibility*. This is the requirement that the chain is able to visit all sets with positive probability under π^* from any starting point in Ω. Formally, a Markov chain is said to be π^*-irreducible if for every $\boldsymbol{x} \in \Omega$,

$$\pi^*(A) > 0 \Rightarrow P(\boldsymbol{\Phi}_i \in A \mid \boldsymbol{\Phi}_0 = \boldsymbol{x}) > 0,$$

for some $i \geqslant 1$. If the space Ω is connected and the function $p(\boldsymbol{x},\boldsymbol{y})$ is positive and continuous, then the Markov chain with transition kernel given by Equation (7) and invariant distribution π^* is π^*-irreducible.

Another important property of a chain is *aperiodicity*, which ensures that the chain does not cycle through a finite number of sets. A Markov chain is aperiodic if there exists no partition of $\Omega = (D_0, D_1, \ldots, D_{p-1})$ for some $p \geqslant 2$ such that $P(\boldsymbol{\Phi}^i \in D_{i\,\mathrm{mod}(p)} \mid \boldsymbol{\Phi}_0 \in D_0) = 1$ for all i.

These definitions allow us to state the following results [see Tierney (1994)], which form the basis for Markov chain Monte Carlo methods. The first of these results gives conditions under which a strong law of large numbers holds and the second gives conditions under which the probability density of the Mth iterate of the Markov chain converges to its unique, invariant density.

Theorem 1. *Suppose $\{\boldsymbol{\Phi}_i\}$ is a π^*-irreducible Markov chain with transition kernel $P(\cdot,\cdot)$ and invariant distribution π^*, then π^* is the unique invariant distribution of $P(\cdot,\cdot)$ and for all π^*-integrable real-valued functions h,*

$$\frac{1}{M}\sum_{i=1}^{M} h(\boldsymbol{\Phi}_i) \to \int h(\boldsymbol{x})\,\pi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \quad \text{as} \quad M \to \infty, \textit{ a.s.}$$

Theorem 2. *Suppose $\{\boldsymbol{\Phi}_i\}$ is a π^*-irreducible, aperiodic Markov chain with transition kernel $P(\cdot,\cdot)$ and invariant distribution π^*. Then for π^*-almost every $\boldsymbol{x} \in \Omega$, and all sets A*

$$\| P^M(\boldsymbol{x}, A) - \pi^*(A) \| \to 0 \quad \text{as} \quad M \to \infty,$$

where $\|\cdot\|$ denotes the total variation distance.

A further strengthening of the conditions is required to obtain a central limit theorem for sample-path averages. A key requirement is that of an ergodic chain, i.e., chains that are irreducible, aperiodic and positive Harris-recurrent [for a definition of the latter, see Tierney (1994)]. In addition, one needs the notion of geometric ergodicity. An ergodic Markov chain with invariant distribution π^* is a geometrically ergodic if there exists a non-negative real-valued function (bounded in expectation under π^*) and a positive constant $r < 1$ such that

$$\| P^M(\boldsymbol{x}, A) - \pi^*(A) \| \leqslant C(\boldsymbol{x})\, r^n,$$

for all $\boldsymbol{x}$ and all n and sets A. Chan and Geyer (1994) show that if the Markov chain is ergodic, has invariant distribution π^*, and is geometrically ergodic, then for all L^2 measurable functions h, taken to be scalar-valued for simplicity, and any initial distribution, the distribution of $\sqrt{M}(\hat{h}_M - \mathrm{E}h)$ converges weakly to a normal distribution with mean zero and variance $\sigma_h^2 \geqslant 0$, where

$$\hat{h}_M = \frac{1}{M}\sum_{i=1}^{M} h(\boldsymbol{\Phi}_i)$$

$$\mathrm{E}h = \int h(\boldsymbol{\Phi})\,\pi(\boldsymbol{\Phi})\,\mathrm{d}\boldsymbol{\Phi},$$

and

$$\sigma_h^2 = \operatorname{Var} h(\boldsymbol{\Phi}_0) + 2\sum_{k=1}^{\infty} \operatorname{Cov}\{h(\boldsymbol{\Phi}_0), h(\boldsymbol{\Phi}_k)\}. \tag{11}$$

3.2. Computation of numerical accuracy and inefficiency factor

Let $\boldsymbol{\Phi}_1, \boldsymbol{\Phi}_2, \ldots, \boldsymbol{\Phi}_M$ denote the output from a Markov chain, possibly collected after discarding the iterates from an initial burn-in period, and suppose that, as above,

$\hat{h}_M = \frac{1}{M}\sum_{i=1}^{M} h(\boldsymbol{\Phi}_i)$ denotes the sample average of the scalar function h. Then, in this context, the variance of $\hat{h}_M$ based on $\{h(\boldsymbol{\Phi}_1), \ldots, h(\boldsymbol{\Phi}_M)\}$ is an estimate of $\sigma_{\hat{h}}^2$ where the square root of the variance of $\hat{h}_M$ is referred to as the *numerical standard error*.

To describe consistent in M estimators of $\sigma_{\hat{h}}^2$, let $Z_i = h(\boldsymbol{\Phi}_i)$ $(i \leqslant M)$. Then, due to the fact that $\{Z_i\}$ is a dependent sequence

$$\begin{aligned}\operatorname{Var}(\hat{h}_M) &= M^{-2}\sum_{j,k}\operatorname{Cov}(Z_j, Z_k)\\ &= s^2M^{-2}\sum_{j,k=1}^{M}\rho_{|j-k|}\\ &= s^2M^{-1}\left\{1 + 2\sum_{s=1}^{M}(1 - \frac{s}{M})\rho_s\right\},\end{aligned}$$

where s^2 is the sample variance of $\{Z_i\}$ and ρ_s is the estimated autocorrelation at lag s [see Ripley (1987, Ch. 6)]. If $\rho_s > 0$ for each s, then this variance is larger than s^2/M which is the variance under independence. Another estimate of the variance can be found by consistently estimating the spectral density f of $\{Z_i\}$ at frequency zero and using the fact that $\operatorname{Var}(\hat{h}_M) = \tau^2/M$, where $\tau^2 = 2\pi f(0)$. Finally, a traditional approach to finding the variance is by the method of "batch means." In this approach, the data $(Z_1, \ldots, Z_M)$ is divided into k batches of length m with means $B_i = m^{-1}[Z_{(i-1)m+1} + \cdots + Z_{im}]$ and the variance of $\hat{h}_M$ estimated as

$$\operatorname{Var}(\hat{h}_M) = \frac{1}{k(k-1)}\sum_{i=1}^{k}(B_i - \bar{B})^2, \tag{12}$$

where the batch size m is chosen to ensure that the first order serial correlation of the batch means is less than 0.05.

Given the numerical variance it is common to calculate the *inefficiency factor,* which is also called the *autocorrelation time,* defined as

$$\kappa_{\hat{h}} = \frac{\operatorname{Var}(\hat{h}_M)}{s^2/M}. \tag{13}$$

This quantity is interpreted as the ratio of the numerical variance of $\hat{h}_M$ to the variance of $\hat{h}_M$ based on independent draws, and its inverse is the relative numerical efficiency defined in Geweke (1992). The inefficiency factor serves to quantify the relative efficiency loss in the computation of $\hat{h}_M$ from correlated versus independent samples.

4. Metropolis–Hastings algorithm

The Metropolis–Hastings (M–H) method is a general MCMC method to produce sample variates from a given multivariate density [Tierney (1994), Chib and Greenberg

(1995a)]. It is based on a candidate generating density that is used to supply a proposal value and a probability of move that is used to determine if the proposal value should be taken as the next item of the chain. The probability of move is based on the ratio of the target density (evaluated at the proposal value in the numerator and the current value in the denominator) times the ratio of the proposal density (at the current value in the numerator and the proposal value in the denominator). Because ratios of the target density are involved, knowledge of the normalizing constant of the target density is not required. There are a number of special cases of this method, each defined either by the form of the proposal density or by the form in which the components of $\boldsymbol{\psi}$ are revised, say in one block or in several blocks. The method is extremely general and powerful, it being possible in principle to view almost any MCMC algorithm, in one way or another, as a variant of the M–H algorithm.

4.1. The algorithm

The goal is to simulate the d-dimensional distribution $\pi^*(\boldsymbol{\psi})$, $\boldsymbol{\psi} \in \Psi \subseteq \Re^d$ that has density $\pi(\boldsymbol{\psi})$ with respect to some dominating measure. To define the algorithm, let $q(\boldsymbol{\psi}, \boldsymbol{\psi}')$ denote the *candidate generating density,* also called a proposal density, that is used to supply a candidate value $\boldsymbol{\psi}'$ given the current value $\boldsymbol{\psi}$, and let $\alpha(\boldsymbol{\psi}, \boldsymbol{\psi}')$ denote the function

$$\alpha(\boldsymbol{\psi}, \boldsymbol{\psi}') = \begin{cases} \min\left[\frac{\pi(\boldsymbol{\psi}')\, q(\boldsymbol{\psi}', \boldsymbol{\psi})}{\pi(\boldsymbol{\psi})\, q(\boldsymbol{\psi}, \boldsymbol{\psi}')}, 1\right] & \text{if} \quad \pi(\boldsymbol{\psi})\, q(\boldsymbol{\psi}, \boldsymbol{\psi}') > 0; \\ 1 & \text{otherwise.} \end{cases} \tag{14}$$

Then, in the M–H algorithm, a candidate value $\boldsymbol{\psi}'$ is drawn from the proposal density and taken to be the next item of the chain with probability $\alpha(\boldsymbol{\psi}, \boldsymbol{\psi}')$. If the proposal value is rejected, then the next sampled value is taken to be the current value. In algorithmic form, the simulated values are obtained by the following recursive procedure.

Algorithm 2: Metropolis–Hastings

```
(1) Specify an initial value $\boldsymbol{\psi}^{(0)}$:
(2) Repeat for $j = 1, 2, \ldots, M$.
    (a) Propose
            $\boldsymbol{\psi}' \sim q(\boldsymbol{\psi}^{(j)})$.
    (b) Let
            $\boldsymbol{\psi}^{(j+1)} = \begin{cases} \boldsymbol{\psi}' & \text{if Unif}(0,1) \leqslant \alpha(\boldsymbol{\psi}^{(j)}, \boldsymbol{\psi}'); \\ \boldsymbol{\psi}^{(j)} & \text{otherwise.} \end{cases}$
(3) Return the values $\{\boldsymbol{\psi}^{(1)}, \boldsymbol{\psi}^{(2)}, \ldots, \boldsymbol{\psi}^{(M)}\}$.
```

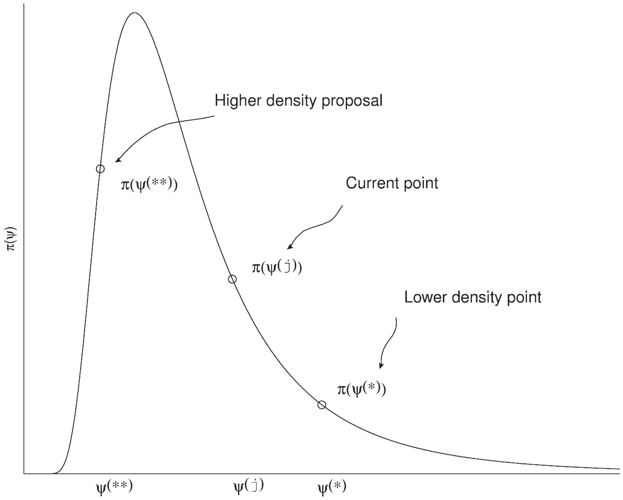

Fig. 2. Original Metropolis algorithm: higher density proposal is accepted with probabability one and the lower density proposal with probability α.

The M–H algorithm delivers variates from π under general conditions. Of course, the variates are from π only in the limit as the number of iterations becomes large but, in practice, after an initial burn-in phase consisting of (say) n_0 iterations, the chain is assumed to have converged and subsequent values are taken as approximate draws from π. Because theoretical calculation of the burn-in is not easy it is important that the proposal density be chosen to ensure that the chain makes large moves through the support of the invariant distribution without staying at one place for many iterations. Generally, the empirical behavior of the M–H output is monitored by the autocorrelation time of each component of ψ and by the *acceptance rate,* which is the proportion of times a move is made as the sampling proceeds.

One should observe that the target density appears as a ratio in the probability $\alpha(\psi, \psi')$ and therefore the algorithm can be implemented without knowledge of the normalizing constant of $\pi(\cdot)$. Furthermore, if the candidate-generating density is symmetric, i.e., $q(\psi, \psi') = q(\psi', \psi)$, the acceptance probability only contains the ratio $\pi(\psi')/\pi(\psi)$; hence, if $\pi(\psi') \geqslant \pi(\psi)$, the chain moves to ψ', otherwise it moves with probability given by $\pi(\psi')/\pi(\psi)$. The latter is the algorithm originally proposed by Metropolis et al. (1953). This version of the algorithm is illustrated in Figure 2.

Different proposal densities give rise to specific versions of the M–H algorithm, each with the correct invariant distribution π. One family of candidate-generating densities is given by $q(\psi, \psi') = q(\psi' - \psi)$. The candidate ψ' is thus drawn according to the process $\psi' = \psi + z$, where z follows the distribution q. Since the candidate is equal to the current value plus noise, this case is called a *random walk M–H* chain. Possible choices for q include the multivariate normal density and the multivariate-t. The random walk M–H chain is perhaps the simplest version of the M–H algorithm [and was the one used by Metropolis et al. (1953)] and quite popular in applications. One has to be careful, however, in setting the variance of z; if it is too large it is possible that the chain may remain stuck at a particular value for many iterations while if it is too small the chain will tend to make small moves and move inefficiently through the support of the target distribution. Both circumstances will tend to generate draws that are highly serially correlated. Note that when q is symmetric, the usual circumstance, $q(z) = q(-z)$ and the probability of move only contains the ratio $\pi(\psi')/\pi(\psi)$. As mentioned earlier, the same reduction occurs if $q(\psi, \psi') = q(\psi', \psi)$.

Hastings (1970) considers a second family of candidate-generating densities that are given by the form $q(\psi, \psi') = q(\psi')$. Tierney (1994) refers to this as an *independence M–H chain* because, in contrast to the random walk chain, the candidates are drawn independently of the current location ψ. In this case, the probability of move becomes

$$\alpha(\psi, \psi') = \min\left\{\frac{w(\psi')}{w(\psi)}, 1\right\},$$

where $w(\psi) = \pi(\psi)/q(\psi)$ is the ratio of the target and proposal densities. For this method to work and not get stuck in the tails of π, it is important that the proposal density have thicker tails than π. A similar requirement is placed on the importance sampling function in the method of importance sampling [Geweke (1989)]. In fact, Mengersen and Tweedie (1996) show that if $w(\psi)$ is uniformly bounded then the resulting Markov chain is ergodic.

Chib and Greenberg (1994) discuss a way of formulating proposal densities in the context of time series autoregressive-moving average models that has a bearing on the choice of proposal density for the independence M–H chain. They suggest matching the proposal density to the target at the mode by a multivariate normal or multivariate-t distribution with location given by the mode of the target and the dispersion given by inverse of the Hessian evaluated at the mode. Specifically, the parameters of the proposal density are taken to be

$$\begin{aligned} &\boldsymbol{m} = \arg\max \log \pi(\psi) \quad \text{and} \\ &\boldsymbol{V} = \tau \left\{ -\frac{\partial^2 \log \pi(\psi)}{\partial\psi\partial\psi'} \right\}^{-1}_{\psi = \hat{\psi}}, \end{aligned} \tag{15}$$

where τ is a tuning parameter that is adjusted to control the acceptance rate. The proposal density is then specified as $q(\psi') = f(\psi'|\boldsymbol{m}, \boldsymbol{V})$, where f is some multivariate density. This may be called a *tailored M–H* chain.

Another way to generate proposal values is through a Markov chain version of the accept–reject method. In this version, due to Tierney (1994), a pseudo accept–reject step is used to generate candidates for an M–H algorithm. Suppose $c > 0$ is a known constant and $h(\psi)$ a source density. Let $C = \{\psi : \pi(\psi) \leqslant ch(\psi)\}$ denote the set of value for which $ch(\psi)$ dominates the target density and assume that this set has high probability under π^*. Now given $\psi^{(n)} = \psi$, the next value $\psi^{(n+1)}$

is obtained as follows: First, a candidate value ψ' is obtained, *independent of the current value* ψ, by applying the accept–reject algorithm with $ch(\cdot)$ as the "pseudo dominating" density. The candidates ψ' that are produced under this scheme have density $q(\psi') \propto \min\{\pi(\psi'), ch(\psi')\}$. If we let $w(\psi) = c^{-1}\pi(\psi)/h(\psi)$ then it can be shown that the M–H probability of move is given by

$$\alpha(\psi, \psi') = \begin{cases} 1 & \text{if } \psi \in C, \\ 1/w(\psi) & \text{if } \psi \notin C, \psi' \in C, \\ \min\{w(\psi')/w(\psi), 1\} & \text{if } \psi \notin C, \psi' \notin C. \end{cases} \tag{16}$$

The choices mentioned above are not exhaustive. Other proposal densities can be generated by mixing over a set of proposal densities, using one proposal density for a certain number of iterations before switching to another.

4.2. Convergence results

In the M–H algorithm the transition kernel of the chain is given by

$$P(\psi, \mathrm{d}\psi') = q(\psi, \psi')\,\alpha(\psi, \psi')\,\mathrm{d}\psi' + r(\psi)\,\delta_\psi(\mathrm{d}\psi'), \tag{17}$$

where $\delta_\psi(\mathrm{d}\psi') = 1$ if $\psi \in \mathrm{d}\psi'$ and 0 otherwise and

$$r(\psi) = 1 - \int_\Omega q(\psi, \psi')\,\alpha(\psi, \psi')\,\mathrm{d}\psi'.$$

Thus, transitions from ψ to ψ' ($\psi' \neq \psi$) are made according to the density

$$p(\psi, \psi') \equiv q(\psi, \psi')\,\alpha(\psi, \psi'), \quad \psi \neq \psi',$$

while transitions from ψ to ψ occur with probability $r(\psi)$. In other words, the density function implied by this transition kernel is of mixed type,

$$K(\psi, \psi') = q(\psi, \psi')\,\alpha(\psi, \psi') + r(\psi)\,\delta_\psi(\psi'), \tag{18}$$

having both a continuous and discrete component where now, with change of notation, $\delta_\psi(\psi')$ is the Dirac delta function defined as $\delta_\psi(\psi') = 0$ for $\psi' \neq \psi$ and $\int_\Omega \delta_\psi(\psi')\,\mathrm{d}\psi' = 1$.

Chib and Greenberg (1995a) provide a way to derive and interpret the probability of move $\alpha(\psi, \psi')$. Consider the proposal density $q(\psi, \psi')$. This proposal density q is not likely to be reversible for π (if it were then we would be done and M–H sampling would not be necessary). Without loss of generality, suppose that $\pi(\psi)\,q(\psi, \psi') > \pi(\psi')\,q(\psi', \psi)$ implying that the rate of transitions from ψ to ψ' exceed those in the reverse direction. To reduce the transitions from ψ to ψ' one can introduce a function $0 \leqslant \alpha(\psi, \psi') \leqslant 1$ such that

$\pi(\boldsymbol{\psi})\, q(\boldsymbol{\psi}, \boldsymbol{\psi}')\, \alpha(\boldsymbol{\psi}, \boldsymbol{\psi}') = \pi(\boldsymbol{\psi}')\, q(\boldsymbol{\psi}', \boldsymbol{\psi})$. Solving for $\alpha(\boldsymbol{\psi}, \boldsymbol{\psi}')$ yields the probability of move in the M–H algorithm. This calculation reveals the important point that the function $p(\boldsymbol{\psi}, \boldsymbol{\psi}') = q(\boldsymbol{\psi}, \boldsymbol{\psi}')\, \alpha(\boldsymbol{\psi}, \boldsymbol{\psi}')$ is reversible by construction, i.e., it satisfies the condition

$$q(\boldsymbol{\psi}, \boldsymbol{\psi}')\, \alpha(\boldsymbol{\psi}, \boldsymbol{\psi}')\, \pi(\boldsymbol{\psi}) = q(\boldsymbol{\psi}', \boldsymbol{\psi})\, \alpha(\boldsymbol{\psi}', \boldsymbol{\psi})\, \pi(\boldsymbol{\psi}'). \tag{19}$$

It immediately follows, therefore, from the argument in Equation (10) that the M–H kernel has $\pi(\boldsymbol{\psi})$ as its invariant density.

It is not difficult to provide conditions under which the Markov chain generated by the M–H algorithm satisfies the conditions of Propositions 1–2. The conditions of Proposition 1 are satisfied by the Metropolis–Hastings chain if $q(\boldsymbol{\psi}, \boldsymbol{\psi}')$ is positive for $(\boldsymbol{\psi}, \boldsymbol{\psi}')$ and continuous and the set $\boldsymbol{\psi}$ is connected. In addition, the conditions of Proposition 2 are satisfied if q is not reversible (which is the usual situation) which leads to a chain that is aperiodic. Conditions for ergodicity, required for use of the central limit theorem, are satisfied if in addition π is bounded. Other similar conditions are provided by Robert and Casella (1999).

4.3. Example

To illustrate the M–H algorithm consider count data taken from Hand et al. (1994) on the number of seizures for 58 epilepsy patients measured first over a eight week baseline period and then over four subsequent two week intervals. At the end of the baseline, each patient is randomly assigned to either a treatment group, which is given the drug Progabide, or a control group which is given a placebo. The model for these data on the ith patient at the jth occasion is taken to be

$$\begin{aligned} y_{ij} | \mathcal{M}, \boldsymbol{\beta} &\sim \text{Poisson}(\lambda_{ij}), \\ \ln(\lambda_{ij}) &= \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \beta_3 x_{3ij} + \ln t_{ij}, \\ \boldsymbol{\beta} &\sim \mathcal{N}_4(0, 10\boldsymbol{I}_4), \end{aligned}$$

where x_1 is an indicator for treatment status, x_2 is an indicator of period, equal to zero for the baseline and one otherwise, $x_3 = x_1 x_2$ and t_{ij} is the offset that is equal to eight in the baseline period and two otherwise. Because the purpose of this example is illustrative, the model does not incorporate the obvious intra-cluster dependence that is likely to be present in the counts.

The target density in this case is the Bayesian posterior density

$$\pi(\boldsymbol{\beta} | \boldsymbol{y}, \mathcal{M}) \propto \pi(\boldsymbol{\beta}) \prod_{i=1}^{58} \prod_{j=0}^{4} \exp(-\lambda_{ij})\, \lambda_{ij}^{y_{ij}},$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)$ and $\pi(\boldsymbol{\beta})$ is the density of the $\mathcal{N}(0, 10\boldsymbol{I}_4)$ distribution. To draw sample variates on $\boldsymbol{\beta}$ from this density we apply the AR–M–H chain.

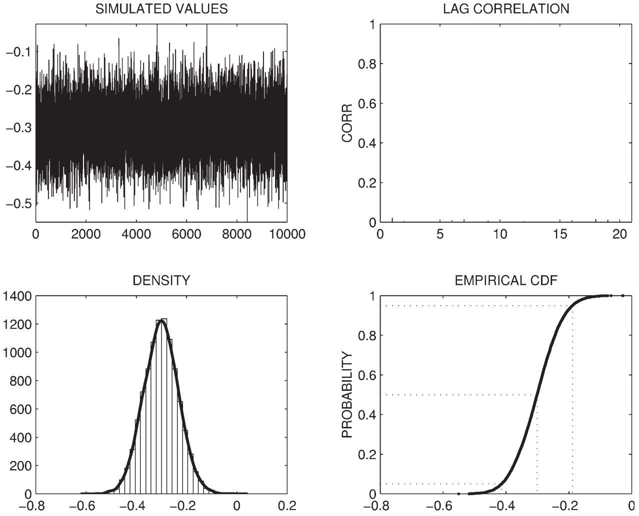

Fig. 3. Marginal posterior distribution of β_1 in Poisson count example. Top left, simulated values by iteration; top right, autocorrelation function of simulated values; bottom left, histogram and superimposed kernel density estimate of marginal density; bottom right, empirical cdf with .05 percentile, 50th percentile and 97.5th percentile marked.

Let $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{V}$ denote the maximum likelihood estimate and inverse of observed information matrix, respectively. Then, the source density $h(\boldsymbol{\beta})$ for the accept–reject method is specified as $f_T(\boldsymbol{\beta}|\hat{\boldsymbol{\beta}}, \boldsymbol{V}, 15)$, a multivariate-$t$ density with fifteen degrees of freedom. The constant c is set equal to 1.5 which implies that the probability of move in Equation (16) is defined in terms of the weight

$$w(\boldsymbol{\beta}|\boldsymbol{y}, \mathcal{M}) = \frac{\pi(\boldsymbol{\beta}) \prod_{i=1}^{58} \prod_{j=0}^{4} \exp(-\lambda_{ij})\lambda_{ij}^{y_{ij}}}{1.5 f_T(\boldsymbol{\beta}|\hat{\boldsymbol{\beta}}, \boldsymbol{V}, 15)}.$$

The MCMC sampler is now run for 10 000 iterations beyond a burn-in of 200 iterations. Of interest in this case is the marginal posterior of β_1 which is summarized in Figure 3.

The figure includes a time series plot of the sampled values, against iteration, and the associated autocorrelation function. These indicate that there is no sign of serial correlation in the sampled values. Although mixing of this kind is often not achieved, this example shows that it is sometimes possible to have a MCMC algorithm produce virtually i.i.d. draws from the target distribution. We also summarize the marginal posterior distribution by a histogram/kernel smoothed plot and the empirical cumulative distribution function. Because the entire distribution is concentrated on negative values it appears that the drug Progabide tends to lower the seizure counts, conditional on the specified model.

4.4. Multiple-block M–H algorithm

In applications when the dimension of ψ is quite large it is preferable to construct the Markov chain simulation by first grouping the variables ψ into p blocks $(\psi_1, \ldots, \psi_p)$, with $\psi_k \in \Omega_k \subseteq \Re^{d_k}$, and sampling each block, conditioned on the rest, by the M–H algorithm. Hastings (1970) considers this general situation and mentions different possibilities for constructing a Markov chain on the product space $\Omega = \Omega_1 \times \cdots \times \Omega_p$.

Let $\psi_{-k} = (\psi_1, \ldots, \psi_{k-1}, \psi_{k+1}, \ldots, \psi_p)$ denote the variables (blocks) excluding ψ_k, in order to describe the multiple-block M–H algorithm. Also let $\pi(\psi_k, \psi_{-k})$ denote the joint density of ψ, regardless of where ψ_k appears in the list $(\psi_1, \ldots, \psi_p)$. Furthermore, let $\{q_k(\psi_k, \psi_k' | \psi_{-k}),\ k \leqslant p\}$ denote a collection of proposal densities, one for each block ψ_k, where the proposal density q_k may depend on the current value of the remaining blocks and is specified along the lines mentioned in connection with the single-block M–H algorithm. Finally, define

$$\alpha_k(\psi_k, \psi_k' | \psi_{-k}) = \min\left\{ \frac{\pi(\psi_k', \psi_{-k})\, q_k(\psi_k', \psi_k | \psi_{-k})}{\pi(\psi_k, \psi_{-k})\, q_k(\psi_k, \psi_k' | \psi_{-k})}, 1 \right\}, \tag{20}$$

as the probability of move for block ψ_k conditioned on ψ_{-k}. Then, in the multiple-block M–H algorithm, one cycle of the algorithm is completed by updating each block, say sequentially in fixed order, using a M–H step with the above probability of move, given the most current value of the remaining blocks. The algorithm may be summarized as follows.

Algorithm 3: Multiple-block Metropolis–Hastings

```
(1) Specify an initial value $\psi^{(0)} = (\psi_1^{(0)}, \ldots, \psi_p^{(0)})$
(2) Repeat for $j = 1, 2, \ldots, M$
    (a) Repeat for $k = 1, 2, \ldots, p$
        (i) Propose
                $\psi_k' \sim q(\psi_k^{(j)}, \psi_k' | \psi_{-k})$.
        (ii) Calculate
                $\alpha_k(\psi_k^{(j)}, \psi_k' | \psi_{-k}) = \min\left\{ \frac{\pi(\psi_k', \psi_{-k})\, q_k(\psi_k', \psi_k^{(j)} | \psi_{-k})}{\pi(\psi_k^{(j)}, \psi_{-k})\, q_k(\psi_k^{(j)}, \psi_k' | \psi_{-k})}, 1 \right\}$.
        (iii) Set
                $\psi_k^{(j+1)} = \begin{cases} \psi_k' & \text{if } \mathrm{Unif}(0,1) \leqslant \alpha_k(\psi_k^{(j)}, \psi_k' | \psi_{-k}) \\ \psi_k^{(j)} & \text{otherwise.} \end{cases}$
(3) Return the values $\{\psi^{(1)}, \psi^{(2)}, \ldots, \psi^{(M)}\}$.
```

Before we examine this algorithm, some features of this method should be noted. First, the version of the algorithm presented above assumes that the blocks are revised sequentially in fixed order. This is not necessary and the blocks may be updated in random order. Second, at the moment block k is updated in this algorithm, the blocks $(\psi_1, \ldots, \psi_{k-1})$ have already been revised while the blocks $(\psi_{k+1}, \ldots, \psi_p)$ have not. Thus, at each step of the algorithm one must be sure to condition on the *most current value of the blocks in* ψ_{-k}. Finally, if the proposal density q_k is determined by tailoring to $\pi(\psi_k, \psi_{-k})$, as in Chib and Greenberg (1994), then this implies that the proposal density is not fixed but varies across iterations.

To understand the multiple-block M–H algorithm, first note that the transition kernel of the kth block, conditioned on ψ_{-k}, may be expressed as

$$P_k(\psi_k, \mathrm{d}\psi_k' | \psi_{-k}) = q(\psi_k, \psi_k' | \psi_{-k})\, \alpha(\psi_k, \psi_k' | \psi_{-k})\, \mathrm{d}\psi_k' + r(\psi_k | \psi_{-k})\, \delta_{\psi_k}(\mathrm{d}\psi_k'), \tag{21}$$

where the notation is similar to that of Equation (17). It can be readily shown that, for a given ψ_{-k}, this kernel satisfies what may be called the *local reversibility condition*

$$\pi(\psi_k | \psi_{-k})\, q(\psi_k, \psi_k' | \psi_{-k})\, \alpha(\psi_k, \psi_k' | \psi_{-k}) = \pi(\psi_k' | \psi_{-k})\, q(\psi_k', \psi_k | \psi_{-k})\, \alpha(\psi_k', \psi_k | \psi_{-k}). \tag{22}$$

As a consequence, the transition kernel of the move from $\psi = (\psi_1, \psi_2, \ldots, \psi_k)$ to $\psi' = (\psi_1', \psi_2', \ldots, \psi_k')$, under the assumption that the blocks are revised sequentially in fixed order, is given by the product of transition kernels

$$P(\psi, \mathrm{d}\psi') = \prod_{k=1}^{p} P_k(\psi_k, \mathrm{d}\psi_k' | \psi_{-k}). \tag{23}$$

This transition kernel is not reversible, as can be easily checked, because under fixed sequential updating of the blocks updating in the reverse order never occurs. The multiple-block M–H algorithm, however, satisfies the weaker condition of invariance. To show this, we follow Chib and Greenberg (1995a). Consider for notational simplicity the case of two blocks, $\psi = (\psi_1, \psi_2)$, where $\psi_k : d_k \times 1$. Now, due to the fact that the local moves satisfy the local reversibility condition (22), the transition kernel $P_1(\psi_1, d\psi_1 | \psi_2)$ has $\pi^*_{1|2}(\cdot | \psi_2)$ as its local invariant distribution (with density $\pi_{1\,|\,2}(\cdot | \psi_2)$), i.e.,

$$\pi^*_{1\,|\,2}(\mathrm{d}\psi_1 | \psi_2) = \int P_1(\psi_1, \mathrm{d}\psi_1 | \psi_2)\, \pi_{1\,|\,2}(\psi_1 | \psi_2)\, \mathrm{d}\psi_1. \tag{24}$$

Similarly, the conditional transition kernel $P_2(\psi_2, \mathrm{d}\psi_2 | \psi_1)$ has $\pi^*_{2\,|\,1}(\cdot|\psi_1)$ as its invariant distribution, for a given value of ψ_1. Then, the kernel formed by multiplying the conditional kernels is invariant for $\pi^*(\cdot,\cdot)$:

$$
\begin{aligned}
&\int\!\!\int P_1(\psi_1, \mathrm{d}\psi_1' | \psi_2) P_2(\psi_2, \mathrm{d}\psi_2' | \psi_1')\, \pi(\psi_1, \psi_2)\, \mathrm{d}\psi_1\, \mathrm{d}\psi_2 \\
&= \int P_2(\psi_2, \mathrm{d}\psi_2' | \psi_1') \left[\int P_1(\psi_1, \mathrm{d}\psi_1' | \psi_2)\, \pi_{1|2}(\psi_1 | \psi_2)\, \mathrm{d}\psi_1 \right] \pi_2(\psi_2)\, \mathrm{d}\psi_2 \\
&= \int P_2(\psi_2, \mathrm{d}\psi_2' | \psi_1')\, \pi^*_{1\,|\,2}(\mathrm{d}\psi_1' | \psi_2)\, \pi_2(\psi_2)\, \mathrm{d}\psi_2 \\
&= \int P_2(\psi_2, \mathrm{d}\psi_2' | \psi_1') \frac{\pi_{2\,|\,1}(\psi_2 | \psi_1')\, \pi_1^*(\mathrm{d}\psi_1')}{\pi_2(\psi_2)} \pi_2(\psi_2)\, \mathrm{d}\psi_2 \\
&= \pi_1^*(\mathrm{d}\psi_1') \int P_2(\psi_2, \mathrm{d}\psi_2' | \psi_1')\, \pi_{2\,|\,1}(\psi_2 | \psi_1')\, \mathrm{d}\psi_2 \\
&= \pi_1^*(\mathrm{d}\psi_1')\, \pi^*_{2\,|\,1}(\mathrm{d}\psi_2' | \psi_1') \\
&= \pi^*(\mathrm{d}\psi_1', \mathrm{d}\psi_2'),
\end{aligned}
$$

where the third line follows from Equation (24), the fourth from Bayes theorem, the sixth from assumed invariance of P_2, and the last from the law of total probability.

The implication of this "product of kernels" result is that it allows us to take draws in succession from each of the kernels, instead of having to run each to convergence for every value of the conditioning variable.

5. The Gibbs sampling algorithm

Another MCMC method, which is a special case of the multiple-block Metropolis–Hastings method, is called the Gibbs sampling method and was brought into statistical prominence by Gelfand and Smith (1990). An elementary introduction to Gibbs sampling is provided by Casella and George (1992). In this algorithm the parameters are grouped into p blocks $(\psi_1, \ldots, \psi_p)$ and each block is sampled according to the *full conditional distribution* of block ψ_k, defined as the conditional distribution under π of ψ_k given all the other blocks ψ_{-k} and denoted as $\pi(\psi_k | \psi_{-k})$. In parallel with the multiple-block M–H algorithm, the most current value of the remaining blocks is used in deriving the full conditional distribution of each block. Derivation of the full conditional distributions is usually quite simple since, by Bayes theorem, $\pi(\psi_k | \psi_{-k}) \propto \pi(\psi_k, \psi_{-k})$, the joint distribution of all the blocks. In addition, the powerful device of data augmentation, due to Tanner and Wong (1987), in which latent or auxiliary variables are artificially introduced into the sampling, is often used to simplify the derivation and sampling of the full conditional distributions.

5.1. The algorithm

To define the Gibbs sampling algorithm, let the set of full conditional distributions be

$$\{\pi(\psi_1|\psi_2, \ldots, \psi_p); \pi(\psi_2|\psi_1, \psi_3, \ldots, \psi_p); \ldots, \pi(\psi_p|\psi_1, \ldots, \psi_{d-1})\}.$$

Now one cycle of the Gibbs sampling algorithm is completed by simulating $\{\psi_k\}_{k=1}^{p}$ from these distributions, recursively updating the conditioning variables as one moves through each distribution. When $d = 2$ one obtains the two block Gibbs sampler that is featured in the work of Tanner and Wong (1987). The Gibbs sampler in which each block is revised in fixed order is defined as follows.

Algorithm 4: Gibbs sampling

```
(1) Specify an initial value $\psi^{(0)} = (\psi_1^{(0)}, \ldots, \psi_p^{(0)})$
(2) Repeat for $j = 1, 2, \ldots, M$
     Generate $\psi_1^{(j+1)}$ from $\pi(\psi_1|\psi_2^{(j)}, \psi_3^{(j)}, \ldots, \psi_p^{(j)})$.
     Generate $\psi_2^{(j+1)}$ from $\pi(\psi_2|\psi_1^{(j+1)}, \psi_3^{(j)}, \ldots, \psi_p^{(j)})$.
     ⋮
     Generate $\psi_p^{(j+1)}$ from $\pi(\psi_p|\psi_1^{(j+1)}, \ldots, \psi_{p-1}^{(j+1)})$.
(3) Return the values $\{\psi^{(1)}, \psi^{(2)}, \ldots, \psi^{(M)}\}$.
```

Thus, the transition of ψ_k from $\psi_k^{(j)}$ to $\psi_k^{(j+1)}$ is effected by taking a draw from the conditional distribution

$$\pi\left(\psi_k|\psi_1^{(j+1)}, \ldots, \psi_{k-1}^{(j+1)}, \psi_{k+1}^{(j)}, \ldots, \psi_p^{(j)}\right),$$

where the conditioning elements reflect the fact that when the kth block is reached, the previous $(k-1)$ blocks have already been updated. The transition density of the chain, again under the maintained assumption that π is absolutely continuous, is therefore given by the product of transition kernels for each block:

$$K\left(\psi^{(j)}, \psi^{(j+1)}\right) = \prod_{k=1}^{p} \pi\left(\psi_k|\psi_1^{(j+1)}, \ldots, \psi_{k-1}^{(j+1)}, \psi_{k+1}^{(j)}, \ldots, \psi_p^{(j)}\right). \tag{25}$$

To illustrate the manner in which the blocks are revised, we consider a two block case, each with a single component, and trace out in Figure 4 a possible trajectory of the sampling algorithm. The contours in the plot represent the joint distribution of ψ and the labels "(0)", "(1)", etc., denote the simulated values. Note that one iteration of the algorithm is completed after both components are revised. Also notice that each component is revised along the direction of the coordinate axes. This feature can be a source of problems if the two components are highly correlated because then

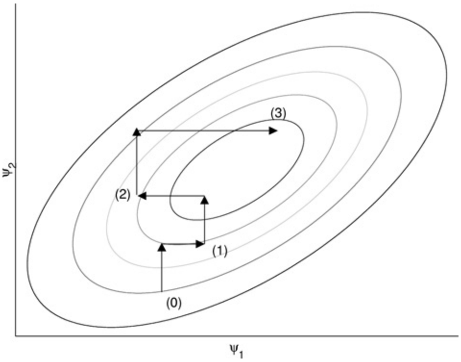

Fig. 4. Gibbs sampling algorithm in two dimensions starting from an initial point and then completing three iterations.

the contours become compressed and movements along the coordinate axes tend to produce only small moves. We return to this issue below.

5.2. Connection with the multiple-block M–H algorithm

A connection with the M–H algorithm can be drawn by noting that the full conditional distribution by Bayes theorem is proportional to the joint distribution, i.e.,

$$\pi(\psi_k|\psi_{-k}) \propto \pi(\psi_k, \psi_{-k}).$$

Now recall that the probability of move in the multiple-block M–H algorithm from Equation (20) is

$$\alpha_k(\psi_k, \psi'_k|\psi_{-k}) = \min\left\{\frac{\pi(\psi'_k, \psi_{-k})\, q(\psi'_k, \psi_k|\psi_{-k})}{\pi(\psi_k, \psi_{-k})\, q(\psi_k, \psi'_k|\psi_{-k})}, 1\right\},$$

so if one substitutes

$$q(\psi_k, \psi'_k|\psi_{-k}) = \pi(\psi'_k, \psi_{-k}),$$
$$q(\psi'_k, \psi_k|\psi_{-k}) = \pi(\psi_k, \psi_{-k}),$$

in this expression all the terms cancel implying that the probability of accepting the proposal is one. Thus, the Gibbs sampling algorithm is a special case of the multiple-block M–H algorithm.

It should be noted that a multiple-block M–H algorithm in which only some of the blocks are sampled using the full conditional distributions are sometimes called *hybrid samplers* or Metropolis-within-Gibbs samplers. These names are not very informative or precise and it is preferable to continue to refer to such algorithms as multiple-block M–H algorithms. The only algorithm that should properly be referred to as the Gibbs algorithm is the one in which each block is sampled directly from its full conditional distribution.

5.3. Invariance of the Gibbs Markov chain

The Gibbs transition kernel is invariant by construction. This is a consequence of the fact that the Gibbs algorithm is a special case of the multiple-block M–H algorithm which is invariant as was established in the last section. A direct calculation also reveals the same result. Consider for simplicity the situation of two blocks when the transition kernel density is

$$K(\psi, \psi') = \pi(\psi_1' \,|\, \psi_2)\, \pi(\psi_2' \,|\, \psi_1').$$

To check invariance we need to show that

$$\int K(\psi, \psi')\pi(\psi_1, \psi_2)\, \mathrm{d}\psi_1 \mathrm{d}\psi_2 = \int \pi(\psi_1' \,|\, \psi_2)\, \pi(\psi_2' \,|\, \psi_1')\, \pi(\psi_1, \psi_2)\, \mathrm{d}\psi_1 \mathrm{d}\psi_2,$$

is equal to $\pi(\psi_1', \psi_2')$. This is easily verified because $\pi(\psi_2' | \psi_1')$ comes out of the integral, and the integral over ψ_1 and ψ_2 produces $\pi(\psi_1')$. This calculation can be extended to any number of blocks in the same way. In addition, the Gibbs Markov chain is not reversible. Reversible Gibbs samplers are discussed by Liu, Wong and Kong (1995).

5.4. Sufficient conditions for convergence

Under rather general conditions, which are easy to verify, the Markov chain generated by the Gibbs sampling algorithm converges to the target density as the number of iterations become large. Formally, if we let $K(\psi, \psi')$ represent the transition density of the Gibbs algorithm and let $K^{(M)}(\psi_0, \psi')$ be the density of the draw ψ' after M iterations given the starting value ψ_0, then

$$\| K^{(M)} \left(\psi^{(0)}, \psi'\right) - \pi(\psi') \| \to 0 \quad \text{as} \quad M \to \infty. \tag{26}$$

Roberts and Smith (1994) [see also Chan (1993)] have shown that the conditions of Proposition 2 are satisfied under the following conditions: (i) $\pi(\psi) > 0$ implies there exists an open neighborhood N_ψ containing ψ and $\epsilon > 0$ such that, for all $\psi' \in N_\psi$, $\pi(\psi') \geqslant \epsilon > 0$; (ii) $\int f(\psi)\, \mathrm{d}\psi_k$ is locally bounded for all k, where ψ_k is the kth block of parameters; and (iii) the support of ψ is arc connected.

It is difficult to find non-pathological problems where these conditions are not satisfied.

5.5. Estimation of density ordinates

We mention that if the full conditional densities are available, whether in the context of the multiple-block M–H algorithm or that of the Gibbs sampler, then the MCMC output can be used to estimate posterior marginal density functions Tanner and Wong (1987) and Gelfand and Smith (1990). One possibility is to use a non-parametric kernel smoothing method which, however, suffers from the curse of dimensionality problem. A more efficient possibility is to exploit the fact that the marginal density of ψ_k at the point ψ_k^* is

$$\pi(\psi_k^*) = \int \pi(\psi_k^* \,|\, \psi_{-k})\, \pi(\psi_{-k}) \mathrm{d}\psi_{-k},$$

where as before $\psi_{-k} = \psi \backslash \psi_k$. Provided the normalizing constant of $\pi(\psi_k^* | \psi_{-k})$ is known, we can estimate the marginal density as an average of the full conditional density over the simulated values of ψ_{-k}:

$$\hat{\pi}(\psi_k^*) = M^{-1} \sum_{j=1}^{M} \pi(\psi_k^* \,|\, \psi_{-k}^{(j)}).$$

Then, under the assumptions of Proposition 1,

$$M^{-1} \sum_{j=1}^{M} \pi(\psi_k^* \,|\, \psi_{-k}^{(j)}) \to \pi(\psi_k^*), \quad \text{as} \quad M \to \infty.$$

Gelfand and Smith (1990) refer to this approach as "Rao–Blackwellization" because of the connections with the Rao–Blackwell theorem in classical statistics. That connection is more clearly seen in the context of estimating (say) the mean of ψ_k, $E(\psi_k) = \int \psi_k \pi(\psi_k)\, \mathrm{d}\psi_k$. By the law of the iterated expectation,

$$E(\psi_k) = E\{E(\psi_k \,|\, \psi_{-k})\},$$

and therefore the estimates

$$M^{-1} \sum_{j=1}^{M} \psi_k^j,$$

and

$$M^{-1} \sum_{j=1}^{M} E(\psi_k \,|\, \psi_{-k}^{(j)}),$$

both converge to $E(\psi_k)$ as $M \to \infty$. Under i.i.d. sampling, and under Markov sampling provided some conditions are satisfied [see Liu, Wong and Kong (1994), Geyer (1995), Casella and Robert (1996) and Robert and Casella (1999)], it can be shown that the variance of the latter estimate is smaller than that of the former. Thus, it can help to average the conditional mean $E(\psi_k | \psi_{-k})$, if that were available, rather than average the draws directly. Gelfand and Smith appeal to this analogy to argue that the Rao–Blackwellized estimate of the density is preferable to that based on the method of kernel smoothing. Chib (1995) extends the Rao–Blackwellization approach to estimate "reduced conditional ordinates" defined as the density of $\boldsymbol{\psi}_k$ conditioned on one or more of the remaining blocks. More discussion of this is provided below in Section 10 on Bayesian model choice. Finally, Chen (1994) provides an importance weighted estimate of the marginal density for cases where the conditional posterior density does not have a known normalizing constant. Chen's estimator is based on the identity

$$\pi(\boldsymbol{\psi}_k^*) = \int w(\boldsymbol{\psi}_k|\boldsymbol{\psi}_{-k}) \frac{\pi(\boldsymbol{\psi}_k^*, \boldsymbol{\psi}_{-k})}{\pi(\boldsymbol{\psi}_k, \boldsymbol{\psi}_{-k})} \pi(\boldsymbol{\psi})\,\mathrm{d}\boldsymbol{\psi},$$

where $w(\boldsymbol{\psi}_k|\boldsymbol{\psi}_{-k})$ is a completely known conditional density whose support is equal to the support of the full conditional density $\pi(\boldsymbol{\psi}_k|\boldsymbol{\psi}_{-k})$. In this form, the normalizing constant of the full conditional density is not required and given a sample of draws $\{\boldsymbol{\psi}^{(1)}, \ldots, \boldsymbol{\psi}^{(M)}\}$ from $\pi(\boldsymbol{\psi})$, a Monte Carlo estimate of the marginal density is given by

$$\hat{\pi}(\boldsymbol{\psi}_k^*) = M^{-1} \sum_{j=1}^{M} w(\boldsymbol{\psi}_k^{(j)}|\boldsymbol{\psi}_{-k}^{(j)}) \frac{\pi(\boldsymbol{\psi}_k^*, \boldsymbol{\psi}_{-k}^{(j)})}{\pi(\boldsymbol{\psi}_k^{(j)}, \boldsymbol{\psi}_{-k}^{(j)})}.$$

Chen (1994) discusses the choice of the conditional density w. Since it depends on $\boldsymbol{\psi}_{-k}$, the choice of w will vary from one sampled draw to the next.

5.6. Example: simulating a truncated multivariate normal

To illustrate the Gibbs sampling algorithm consider the question of sampling a trivariate normal distribution truncated to the positive orthant. In particular, let the target distribution be

$$\pi(\boldsymbol{\psi}) = \frac{1}{\Pr(\boldsymbol{\psi} \in A)} f_N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) I(\boldsymbol{\psi} \in A) \propto f_N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) I(\boldsymbol{\psi} \in A),$$

where $\boldsymbol{\mu} = (.5, 1, 1.5)'$, $\boldsymbol{\Sigma}$ is in equi-correlated form with units on the diagonal and 0.7 on the off-diagonal, $A = (0, \infty) \times (0, \infty) \times (0, \infty)$ and $\Pr(\boldsymbol{\psi} \in A)$ is the normalizing constant which is difficult to compute. Following Geweke (1991), one may define the Gibbs sampler with the blocks ψ_1, ψ_2, ψ_3 and the full conditional distributions

$$\pi(\psi_1|\psi_2, \psi_3);\ \pi(\psi_2|\psi_1, \psi_3);\ \pi(\psi_3|\psi_1, \psi_2),$$

where each of the these full conditional distributions is univariate truncated normal restricted to the interval $(0, \infty)$:

$$\pi(\psi_k|\boldsymbol{\psi}_{-k}) \propto f_N\left(\psi_k|\mu_k + \boldsymbol{C}_k'\boldsymbol{\Sigma}_{-k}^{-1}(\boldsymbol{\psi}_{-k} - \boldsymbol{\mu}_{-k}), \Sigma_k - \boldsymbol{C}_k'\boldsymbol{\Sigma}_{-k}^{-1}\boldsymbol{C}_k\right) I(\psi_k \in (0, \infty)). \tag{27}$$

In this expression we have utilized the well known result about conditional normal distributions and have let $\boldsymbol{C}_k = \mathrm{Cov}(\psi_k, \boldsymbol{\psi}_{-k})$, $\boldsymbol{\Sigma}_{-k} = \mathrm{Var}(\boldsymbol{\psi}_{-k})$ and $\boldsymbol{\mu}_{-k} = E(\boldsymbol{\psi}_{-k})$. Note

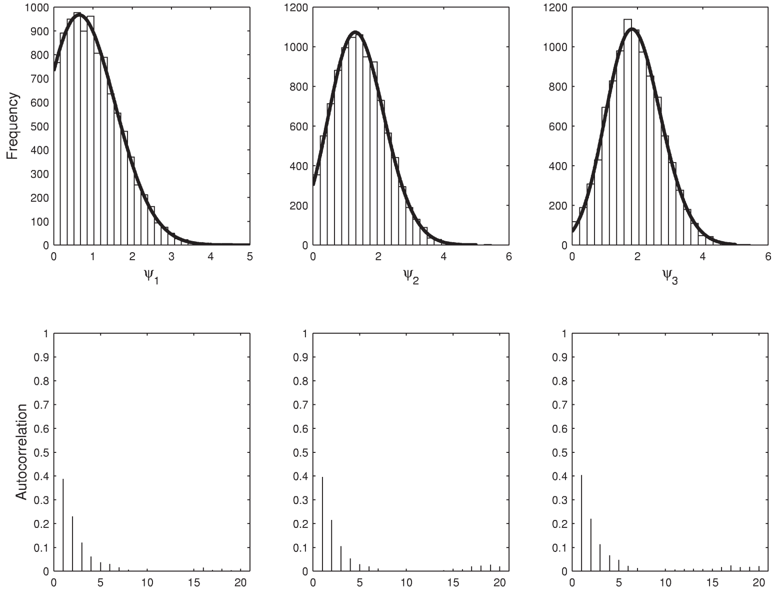

Fig. 5. Marginal distributions of ψ in truncated multivariate normal example (top panel). Histograms of the sampled values and Rao–Blackwellized estimates of the densities are shown. Autocorrelation plots of the Gibbs MCMC chain are in the bottom panel. Graphs are based on 10 000 iterations following a burn-in of 500 cycles.

that, unfortunately, the use of singleton block sizes is unavoidable in this problem because the conditional distribution of any two components given the third is not easy to simulate.

Figure 5 gives the marginal distribution of each component of ψ_k from a Gibbs sampling run of $M = 10\,000$ iterations with a burn-in of 100 cycles. The figure includes both the histograms of the sampled values and the Rao–Blackwellized estimates of the marginal densities based on the averaging of Equation (27) over the simulated values of $\boldsymbol{\psi}_{-k}$. The agreement between the two density estimates is close. In the bottom panel of Figure 5 we plot the autocorrelation function of the sampled draws. The rapid decline in the autocorrelations for higher lags indicates that the sampler is mixing well.

6. Sampler performance and diagnostics

In implementing a MCMC method it is important to assess the performance of the sampling algorithm to determine the rate of mixing and the size of the burn-in, both having implications for the number of iterations required to get reliable answers. A large literature has now emerged on these issues, for example, Robert (1995), Tanner (1996, Section 6.3), Cowles and Carlin (1996), Gammerman (1997, Section 5.4), Brooks, Dellaportas and Roberts (1997) and Robert and Casella (1999), but the ideas, although related in many ways, have not coalesced into a single prescription.

One approach for determining sampler performance and the size of the burn-in time is to employ analytical methods to the specified Markov chain, prior to sampling. This approach is exemplified in the work of, for example, Meyn and Tweedie (1994), Polson (1996), Roberts and Tweedie (1996) and Rosenthal (1995). Two factors have inhibited the growth and application of these methods. The first is that the calculations are difficult and problem-specific, and second, the upper bounds for the burn-in that emerge from such calculations are usually highly conservative.

At this time the more popular approach is to utilize the sampled draws to assess both the performance of the algorithm and its approach to the stationary, invariant distribution. Several such relatively informal methods are now available. Gelfand and Smith (1990) recommend monitoring the evolution of the quantiles as the sampling proceeds. Another quite useful diagnostic, one that is perhaps the simplest and most direct, are autocorrelation plots (and autocorrelation times) of the sampled output. Slowly decaying correlations indicate problems with the mixing of the chain. It is also useful in connection with M–H Markov chains to monitor the acceptance rate of the proposal values with low rates implying "stickiness" in the sampled values and thus a slower approach to the invariant distribution.

Somewhat more formal sample-based diagnostics are also available in the literature, as summarized in the CODA routines provided by Best, Cowles and Vines (1995). Although these diagnostics often go under the name "convergence diagnostics" they are in principle approaches that detect *lack* of convergence. Detection of convergence based entirely on the sampled output, without analysis of the target distribution, is extremely difficult and perhaps impossible. Cowles and Carlin (1996) discuss and evaluate thirteen such diagnostics [for example, those proposed by Geweke (1992), Raftery and Lewis (1992), Ritter and Tanner (1992), Gelman and Rubin (1992), Zellner and Min (1995), amongst others] without arriving at a consensus. Difficulties in evaluating these methods stem from the fact that some of these methods apply only to Gibbs Markov chains [for example, those of Ritter and Tanner (1992) and Zellner and Min (1995)] while others are based on the output not just of a single chain but on that of multiple chains specifically run from "disparate starting values" as in the method of Gelman and Rubin (1992). Finally, some methods assess the behavior of univariate moment estimates [as in the approach of Geweke (1992) and Gelman and Rubin (1992)] while others are concerned with the behavior of the entire transition kernel [as in Ritter and Tanner (1992) and Zellner and Min (1995)]. Further developments in this area are ongoing.

7. Strategies for improving mixing

In practice, while implementing MCMC methods it is important to construct samplers that mix well, where mixing is measured by the autocorrelation time, because such samplers can be expected to converge more quickly to the invariant distribution. Over the years a number of different recipes for designing samplers with low autocorrelation times have been proposed although it may sometimes be difficult, because of the complexity of the problem, to apply any of these recipes.

7.1. *Choice of blocking*

As a general rule, sets of parameters that are highly correlated should be treated as one block when applying the multiple-block M–H algorithm. Otherwise, it would be difficult to develop proposal densities that lead to large moves through the support of the target distribution and the sampled draws would tend to display autocorrelations that decay slowly. To get a sense of the problem, it may be worthwhile for the reader to use the Gibbs sampler to simulate a bivariate normal distribution with unit variances and covariance (correlation) of 0.95.

The importance of coarse, or highly grouped, blocking has been highlighted in a number of different problems for example, the state space model, hidden Markov model and longitudinal data models with random effects. In each of these situations, which are further discussed below in detail, the parameter space is quite large on account of the fact that auxiliary variables are included in the sampling (the latent states in the case of the state space model and the random effects in the case of the longitudinal data model). These latent variables tend to be highly correlated either amongst themselves, as in the case of the state space model, or with a different set of variables as in the case of the panel model.

Blocks can be combined by the method of composition. For example, suppose that ψ_1, ψ_2 and ψ_3 denote three blocks and that the distribution $\psi_1|\psi_3$ is tractable (i.e., can be sampled directly). Then, the blocks (ψ_1, ψ_2) can be collapsed by first sampling ψ_1 from $\psi_1|\psi_3$ followed by ψ_2 from $\psi_2|\psi_1, \psi_3$. This amounts to a two block MCMC algorithm. In addition, if it is possible to sample (ψ_1, ψ_2) marginalized over ψ_3 then the number of blocks is reduced to one. Liu (1994) and Liu, Wong and Kong (1994) discuss the value of these strategies in the context of a three-block Gibbs MCMC chains. Roberts and Sahu (1997) provide further discussion of the role of blocking in the context of Gibbs Markov chains used to sample multivariate normal target distributions.

7.2. *Tuning the proposal density*

As mentioned above, the proposal density in a M–H algorithm has an important bearing on the mixing of the MCMC chain. Fortunately, one has great flexibility in the choice of candidate generating density and it is possible to adapt the choice to the specific context of a given problem. For example, Chib, Greenberg and Winkelmann (1998) develop and compare four different choices in the context of longitudinal random effects for count data. In this problem, each cluster (or individual) has its own random effects and each of these has to be sampled from an intractable target distribution.

If one lets n denote the number of clusters, where n is typically large, say in excess of a thousand, then the number of blocks in the MCMC implementation is $n + 3$ (n for each of the random effect distributions, two for the fixed effects and one for the variance components matrix). For this problem, the multiple-block M–H algorithm requires $n + 1$ M–H steps within one iteration of the algorithm. Tailored proposal densities are therefore computationally quite expensive but one can use a mixture of proposal densities where a less demanding proposal, for example a random walk proposal, is combined with the tailored proposal to sample each of the n random effect target distributions. Further discussion of mixture proposal densities for the purpose of improving mixing is contained in Tierney (1994).

7.3. *Other strategies*

In some problems it is possible to reparameterize the variables to make the blocks less correlated. See Hills and Smith (1992) and Gelfand, Sahu and Carlin (1995) where under certain circumstances reparameterization is shown to be beneficial for simple one-way analysis of variance models, and for general hierarchical normal linear models.

Another strategy that can prove useful is importance resampling in which the MCMC sampler is applied not to the target distribution π but to a modified distribution π^*, for which a well mixing sampler can be designed, and which is close to π. Now suppose $\{\psi^{(1)}, \ldots, \psi^{(M)}\}$ are draws from the target distribution π^*. These can be made to correspond to the target distribution π by attaching the weight $w_j = \pi(\psi^{(j)})/\pi^*(\psi^{(j)})$ to each draw and then re-sampling the sampled values with probability given by $\{w_j / \sum_{g=1}^{M} w_g\}$. This strategy was introduced for a different purpose by Rubin (1988) and then employed by Gelfand and Smith (1992) and Albert (1993) to study the sensitivity of the posterior distribution to small changes in the prior without involving a new MCMC calculation. Its use for improving mixing in the MCMC context is illustrated by Kim, Shephard and Chib (1998) where a nonlinear state space model of stochastic volatility is approximated accurately by a mixture of state space models; an efficient MCMC algorithm is then developed for the latter target distribution and the draws are finally re-sampled to correspond to the original nonlinear model.

Other approaches have also been discussed in the literature. Marinari and Parisi (1992) develop the simulated tempering method whereas Geyer and Thompson (1995) develop a related technique that they call the Metropolis-coupled MCMC method. Both these approaches rely on a series of transition kernels $\{K_1, \ldots, K_m\}$ where only K_1 has π^* as the stationary distribution. The other kernels have equilibrium distributions π_i, which Geyer and Thompson take to be $\pi_i(\psi) = \pi(\psi)^{1/i}$, $i = 2, \ldots, m$. This specification produces a set of target distributions that have higher variance than π^*. Once the transition kernels and equilibrium distributions are specified then the Metropolis-coupled MCMC method requires that each of the m kernels be used in parallel. At each iteration, after the m draws have been obtained, one randomly selects two chains to see if the states should be swapped. The probability of swap is based on the M–H acceptance condition. At the conclusion of the sampling, inference is based on the sequence of draws that correspond to the distribution π^*. These methods promote rapid mixing because draws from the various "flatter" target densities have a chance of being swapped with the draws from the base kernel K_1. Thus, variates that are unlikely under the transition K_1 have a chance of being included in the chain, leading to more rapid exploration of the parameter space.

8. MCMC algorithms in Bayesian estimation

8.1. Overview

Markov chain Monte Carlo methods have proved enormously popular in Bayesian statistics [for wide-ranging discussions of the Bayesian paradigm see, for example, Zellner (1971), Leamer (1978), Berger (1985), O'Hagan (1994), Bernardo and Smith (1994), Poirier (1995), Gelman, Meng, Stern and Rubin (1995)], where these methods have opened up vistas that were unimaginable fifteen years ago. Within the Bayesian framework, where both parameters and data are treated as random variables and inferences about the parameters are conducted conditioned on the data, the posterior distribution of the parameters provides a natural target for MCMC methods. Sometimes the target distribution is the posterior distribution of the parameters augmented by latent data, in which case the MCMC scheme operates on a space that is considerably larger than the parameter space. This strategy, which goes under the name of data augmentation, is illustrated in several models below and its main virtue is that it allows one to conduct the MCMC simulation without having to evaluate the likelihood function of the parameters. The latter feature is of considerable importance especially when the model of interest has a complicated likelihood function and likelihood based inference is difficult. Admittedly, in standard problems such as the linear regression model, there may be little to be gained by utilizing MCMC methods or in fact by adopting the Bayesian approach, but the important point is that MCMC methods provide a complete computational toolkit for conducting Bayesian inference in models that are both simple and complicated. This is the central reason for the current growing appeal of Bayesian methods in theoretical and practical work and this appeal is likely to increase once MCMC Bayesian software, presently under development at various sites, becomes readily available.

Papers that develop some of the important general MCMC ideas for Bayesian inference appeared early in the 1990's. Categorized by topics, these include, normal and student-t data models [Gelfand et al. (1990), Carlin and Polson (1991)]; binary and ordinal response models [Albert and Chib (1993a, 1995)]; tobit censored regression models [Chib (1992)]; generalized linear models [Dellaportas and Smith (1993), Mallick and Gelfand (1994)]; change point models [Carlin et al. (1992), Stephens (1994)]; autoregressive models [Chib (1993), McCulloch and Tsay (1994)];

autoregressive-moving average models [Chib and Greenberg (1994)]; hidden Markov models [Albert and Chib (1993b), Robert et al. (1993), McCulloch and Tsay (1994), Chib (1996)]; state space models [Carlin, Polson and Stoffer (1992), Carter and Kohn (1994, 1996), Chib and Greenberg (1995b), de Jong and Shephard (1995)]; measurement error models [Mallick and Gelfand (1996)]; mixture models [Diebolt and Robert (1994), Escobar and West (1995), Muller, Erkanli and West (1996)]; longitudinal data models [Zeger and Karim (1991), Wakefield et al. (1994)].

More recently, other model and inference situations have also come under scrutiny. Examples include, ARMA models with switching [Billio, Monfort and Robert (1999)]; CART models [Chipman, George and McCulloch (1998), Denison, Mallick and Smith (1998)]; conditionally independent hierarchical models [Albert and Chib (1997)]; estimation of HPD intervals [Chen and Shao (1999)]; item response models [Patz and Junker (1999)]; selection models [Chib and Hamilton (2000)]; partially linear and additive regression models [Lenk (1999), Shively, Kohn and Wood (1999)]; sequential Monte Carlo for state space models [Liu and Chen (1998), Pitt and Shephard (1999)]; stochastic differential equation models [Elerian, Chib and Shephard (1999)]; models with symmetric stable distributions [Tsionas (1999)]; neural network models [Muller and Insua (1998)]; spatial models [Waller, Carlin, Xia and Gelfand (1997)].

MCMC methods have also been extended to the realm of Bayesian model choice. Problems related to variable selection in regression models, hypothesis testing in nested models and the general problem of model choice are now all amenable to analysis by MCMC methods. The basic strategies are developed in the following papers: variable selection in regression [George and McCulloch (1993)]; hypothesis testing in nested models [Verdinelli and Wasserman (1995)]; predictive model comparison [Gelfand and Dey (1994)]; marginal likelihood and Bayes factor computation [Chib (1995)]; composite model space and parameter space MCMC [Carlin and Chib (1995), Green (1995)]. These developments are discussed in Section 10.

We now provide a set of applications of MCMC methods to models largely drawn from the list above. These models serve to illustrate a number of general techniques, for example, derivations of full conditional distributions, use of latent variables in the sampling (data augmentation) to avoid computation of the likelihood function, and issues related to blocking. Because of the modular nature of MCMC methods, the algorithms presented below can serve as the building blocks for other models not considered here. In some instances one would only need to combine different pieces of these algorithms to fit a new model.

8.2. Notation and assumptions

To streamline the discussion we collect some of the notation that is used in the rest of the paper.

The d-variate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Omega}$ is denoted by $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Omega})$. Its density at the point $\boldsymbol{t} \in \Re^d$ is denoted by $\phi_d(\boldsymbol{t}|\boldsymbol{\mu}, \boldsymbol{\Omega})$. The univariate normal density truncated to the interval (a, b) is denoted by $\mathcal{TN}_{[a,b]}(\mu, \sigma^2)$

with density at the point $t \in (a, b)$ given by $\phi(t|\mu, \sigma^2)/[\Phi((b-\mu)/\sigma) - \Phi((a-\mu)/\sigma)]$, where ϕ is the univariate normal density and $\Phi(\cdot)$ is the c.d.f. of the standard normal random variable.

A d-variate random vector distributed according to the multivariate-t distribution with mean vector $\boldsymbol{\mu}$, dispersion matrix $\boldsymbol{\Sigma}$ and ξ degrees of freedom has density $f_T(\boldsymbol{t}|\boldsymbol{\mu}, \boldsymbol{\Omega}, \xi)$ given by

$$\frac{\Gamma((\xi+1)/2)\Gamma(\xi/2)}{(\xi\pi)^{1/2}|\boldsymbol{\Omega}|^{1/2}} \left\{1 + \frac{1}{\xi}(\boldsymbol{t}-\boldsymbol{\mu})'\boldsymbol{\Omega}^{-1}(\boldsymbol{t}-\boldsymbol{\mu})\right\}^{-(\xi+d)/2}.$$

The gamma distribution is denoted by $\Gamma(a, b)$ with density at the point t by $f_G(t|a, b) \propto t^{a-1}\exp(-bt)I[t > 0]$, where $I[A]$ is the indicator function of the event A. The inverse gamma distribution is the distribution of the inverse of a gamma variate.

A random symmetric positive definite matrix $\boldsymbol{W}: p \times p$ is said to follow a Wishart distribution $\mathcal{W}_p(\boldsymbol{W}|\upsilon, \boldsymbol{R})$ if the density of $\boldsymbol{W}$ is given by

$$c\frac{|\boldsymbol{W}|^{(v-p-1)/2}}{|\boldsymbol{R}|^{v/2}} \exp\left\{-\tfrac{1}{2}\operatorname{tr}(\boldsymbol{R}^{-1}\boldsymbol{W})\right\}, \quad |\boldsymbol{W}| > 0,$$

where c is a normalizing constant, $\boldsymbol{R}$ is a hyperparameter matrix and "tr" is the trace function. To simulate the Wishart distribution, one utilizes the expression $\boldsymbol{W} = \boldsymbol{LTT'L'}$, where $\boldsymbol{R} = \boldsymbol{LL'}$ and $\boldsymbol{T} = (t_{ij})$ is a lower triangular matrix with $t_{ii} \sim \sqrt{\chi^2_{\upsilon-i+1}}$ and $t_{ij} \sim \mathcal{N}(0, 1)$.

In connection with the sampling design of the observations and the error terms we use "ind" to denote independent and "i.i.d." to denote independent and identically distributed. The response variable (or vector) of the model is denoted by either y_i or y_t, the sample size by n and the entire collection of sample data by $\boldsymbol{y} = (y_1, \ldots, y_n)$. In some instances, we let $\boldsymbol{Y}_t = (y_1, \ldots, y_t)$ denote the data upto time t and $\boldsymbol{Y}^t = (y_t, \ldots, y_n)$ to denote the values from t to the end of the sample. The covariates are denoted as x_i if the corresponding response is a scalar and as $\boldsymbol{X}_i$ or $\boldsymbol{X}_t$ if the response is a vector. The regression coefficients are denoted by $\boldsymbol{\beta}$ and the error variance (if y_i is a scalar) by σ^2 and the error covariance by $\boldsymbol{\Omega}$ if y_i is a vector. The parameters of the model are denoted by $\boldsymbol{\theta}$ and the variables used in the MCMC simulation by $\boldsymbol{\psi}$ (consisting of $\boldsymbol{\theta}$ and other quantities).

When denoting conditional distributions only dependence on random quantities, such as parameters and random effects, is included in the conditioning set. Covariates are never included in the conditioning. The symbol p is used to denote the prior density if general notation is required.

It is always assumed that each distinct set of parameters, for example, regression coefficients and covariance elements, are a priori independent. The joint prior distribution is therefore specified through the marginal distribution of each distinct set of parameters. Distributions for the parameters are chosen from the class of conditionally conjugate distributions in keeping with the existing literature on these models. The parameters of the prior distributions, called hyperparameters, are assumed known. These will be indicated by the subscript "0." In some cases, when the hyperparameters are unknown, hierarchical priors, defined by placing prior distributions on the prior hyperparameters, are used.

8.3. Normal and student-t regression models

Consider the univariate regression model defined by the specification

$$\begin{aligned} y_i|\mathcal{M},\boldsymbol{\beta},\sigma^2 &\sim \mathcal{N}(\boldsymbol{x}_i'\boldsymbol{\beta},\sigma^2), \quad i \leqslant n,\\ \boldsymbol{\beta} &\sim \mathcal{N}_k(\boldsymbol{\beta}_0,\boldsymbol{B}_0),\\ \sigma^2 &\sim \mathcal{IG}\left(\frac{\upsilon_0}{2},\frac{\delta_0}{2}\right). \end{aligned}$$

The target distribution is

$$\pi(\boldsymbol{\beta},\sigma^2|\mathcal{M},\boldsymbol{y}) \propto p(\boldsymbol{\beta})p(\sigma^2)\prod_{i=1}^{n} f(y_i|\boldsymbol{x}_i'\boldsymbol{\beta},\sigma^2),$$

and MCMC simulation proceeds by a Gibbs chain defined through the full conditional distributions

$$\boldsymbol{\beta}|\boldsymbol{y},\mathcal{M},\sigma^2;\ \sigma^2|\boldsymbol{y},\mathcal{M},\boldsymbol{\beta}.$$

Each of these distributions is straightforward to derive because conditioned on σ^2 both the prior and the likelihood have Gaussian forms (and hence the updated distribution is Gaussian with moments found by completing the square for the terms in the exponential function) while conditioned on $\boldsymbol{\beta}$, the updated distribution of σ^2 is inverse gamma with parameters found by adding the exponents of the prior and the likelihood.

Algorithm 5: Gaussian multiple regression

(1) `Sample`

$$\boldsymbol{\beta} \sim \mathcal{N}_k\left(\boldsymbol{B}_n\left(\boldsymbol{B}_0^{-1}\boldsymbol{\beta}_0+\sigma^{-2}\sum_{i=1}^{n}\boldsymbol{x}_i y_i\right), \boldsymbol{B}_n = \left(\boldsymbol{B}_0^{-1}+\sigma^{-2}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}\right)$$

(2) `Sample`

$$\sigma^2 \sim \mathcal{IG}\left\{\frac{\upsilon_0+n}{2},\frac{\delta_0+\sum_{i=1}^{n}(y_i-\boldsymbol{x}_i'\boldsymbol{\beta})^2}{2}\right\}$$

(3) `Goto 1.`

This algorithm can be easily modified to permit the observations y_i to follow a Student-*t* distribution. The modification, proposed by Carlin and Polson (1991), utilizes the fact that if

$$\lambda_i \sim \mathcal{G}\left(\frac{\xi}{2}, \frac{\xi}{2}\right),$$

and

$$y_i|\mathcal{M}, \boldsymbol{\beta}, \sigma^2, \lambda_i \sim \mathcal{N}(\boldsymbol{x}_i'\boldsymbol{\beta}, \lambda_i^{-1}\sigma^2),$$

then

$$y_i|\mathcal{M}, \boldsymbol{\beta}, \sigma^2 \sim f_T(y_i|\boldsymbol{x}_i'\boldsymbol{\beta}, \sigma^2, \xi), \quad i \leqslant n.$$

Hence, if one defines $\boldsymbol{\psi} = (\boldsymbol{\beta}, \sigma^2, \{\lambda_i\})$ then, conditioned on $\{\lambda_i\}$, the model is Gaussian and a variant of Algorithm 5 can be used. Furthermore, conditioned on $(\boldsymbol{\beta}, \sigma^2)$, the full conditional distribution of $\{\lambda_i\}$ factors into a product of independent Gamma distributions.

Algorithm 6: Student-*t* multiple regression

(1) `Sample`

$$\boldsymbol{\beta} \sim \mathcal{N}_k\left(\boldsymbol{B}_{n,\lambda}\left(\boldsymbol{B}_0^{-1}\boldsymbol{\beta}_0 + \sigma^{-2}\sum_{i=1}^{n}\lambda_i\boldsymbol{x}_i y_i\right), \boldsymbol{B}_{n,\lambda} = \left(\boldsymbol{B}_0^{-1} + \sigma^{-2}\sum_{i=1}^{n}\lambda_i\boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}\right).$$

(2) `Sample`

$$\sigma^2 \sim \mathcal{IG}\left\{\frac{\upsilon_0 + n}{2}, \frac{\delta_0 + \sum_{i=1}^{n}\lambda_i(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2}{2}\right\}.$$

(3) `Sample`

$$\lambda_i \sim \mathcal{G}\left[\frac{\xi + 1}{2}, \frac{\xi + \sigma^{-2}(y_i - \boldsymbol{x}_i\boldsymbol{\beta})^2}{2}\right], \quad i \leqslant n.$$

(4) `Goto 1.`

Another modification of Algorithm 5 is to Zellner's seemingly unrelated regression model (SUR). In this case a vector of p observations are generated from the model

$$\begin{aligned} y_t|\mathcal{M}, \boldsymbol{\beta}, \boldsymbol{\Omega} &\sim \mathcal{N}(\boldsymbol{X}_t\boldsymbol{\beta}, \boldsymbol{\Omega}), \quad t \leqslant n, \\ \boldsymbol{\beta} &\sim \mathcal{N}_k(\boldsymbol{\beta}_0, \boldsymbol{B}_0), \\ \boldsymbol{\Omega}^{-1} &\sim \mathcal{W}_p(\nu_0, \boldsymbol{R}_0), \end{aligned}$$

where $\boldsymbol{y}_t = (y_{1t}, \ldots, y_{pt})'$, $\boldsymbol{X}_t = \operatorname{diag}(\boldsymbol{x}_{1t}', \ldots, \boldsymbol{x}_{pt}')$, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_p')' : k \times 1$, and $k = \sum_i k_i$.

To deal with this model, a two block MCMC approach can be used as proposed by Blattberg and George (1991) and Percy (1992). Chib and Greenberg (1995b) extend that algorithm to SUR models with hierarchical priors and time-varying parameters of the type considered by Gammerman and Migon (1993).

For the SUR model, the posterior density of the parameters is proportional to

$$\pi(\boldsymbol{\beta})\pi(\boldsymbol{\Omega}^{-1}) \times \left|\boldsymbol{\Omega}^{-1}\right|^{n/2} \exp\left\{-\tfrac{1}{2}\sum_{t=1}^{n}(\boldsymbol{y}_t-\boldsymbol{X}_t\boldsymbol{\beta})'\boldsymbol{\Omega}^{-1}(\boldsymbol{y}_t-\boldsymbol{X}_t\boldsymbol{\beta})\right\},$$

and the MCMC algorithm is defined by the full conditional distributions

$$\boldsymbol{\beta}|\boldsymbol{y},\mathcal{M},\boldsymbol{\Omega}^{-1};\ \boldsymbol{\Omega}^{-1}|\boldsymbol{y},\mathcal{M},\boldsymbol{\beta}.$$

These are both tractable, with the former a normal distribution and the latter a Wishart distribution.

Algorithm 7: Gaussian SUR

(1) `Sample`

$$\boldsymbol{\beta} \sim \mathcal{N}_k\left(\boldsymbol{B}_n\left(\boldsymbol{B}_0^{-1}\boldsymbol{\beta}_0+\sum_{t=1}^{n}\boldsymbol{X}_t'\boldsymbol{\Omega}^{-1}\boldsymbol{y}_t\right), \boldsymbol{B}_n=\left(\boldsymbol{B}_0^{-1}+\sum_{t=1}^{n}\boldsymbol{X}_t'\boldsymbol{\Omega}^{-1}\boldsymbol{X}_t\right)^{-1}\right).$$

(2) `Sample`

$$\boldsymbol{\Omega}^{-1} \sim \mathcal{W}_p\left[\nu_0+n, \left\{\boldsymbol{R}_0^{-1}+\sum_{t=1}^{n}(\boldsymbol{y}_t-\boldsymbol{X}_t\boldsymbol{\beta})(\boldsymbol{y}_t-\boldsymbol{X}_t\boldsymbol{\beta})'\right\}^{-1}\right].$$

(3) `Goto 1.`

8.4. Binary and ordinal probit

Suppose that each y_i is binary and the model of interest is

$$y_i|\mathcal{M},\boldsymbol{\beta} \sim \Phi(\boldsymbol{x}_i'\boldsymbol{\beta}), i \leqslant n; \quad \boldsymbol{\beta} \sim \mathcal{N}_k(\boldsymbol{\beta}_0,\boldsymbol{B}_0).$$

The posterior distribution does not belong to a named family of distributions. To deal with the problem, Albert and Chib (1993a) introduce a technique that has formed the basis for a unified methodology for univariate and multivariate binary and ordinal response models and led to many applications. The Albert–Chib algorithm capitalizes on the simplifications afforded by introducing latent or auxiliary data into the sampling.

Instead of the specification above, the model of interest is specified in equivalent form as

$$z_i|\mathcal{M}, \boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{x}_i'\boldsymbol{\beta}, 1), \quad y_i = I[z_i > 0], i \leqslant n, \quad \boldsymbol{\beta} \sim \mathcal{N}_k(\boldsymbol{\beta}_0, \boldsymbol{B}_0).$$

Now the MCMC Gibbs algorithm proceeds with the sampling of the full conditional distributions

$$\boldsymbol{\beta}|\boldsymbol{y}, \mathcal{M}, \{z_i\}; \quad \{z_i\}|\boldsymbol{y}, \mathcal{M}, \boldsymbol{\beta},$$

where

$$\boldsymbol{\beta}|\boldsymbol{y}, \mathcal{M}, \{z_i\} \overset{d}{=} \boldsymbol{\beta}|\mathcal{M}, \{z_i\},$$

has the same form as in the linear regression model with σ^2 set equal to one and y_i replaced by z_i and

$$\{z_i\}|\boldsymbol{y}, \mathcal{M}, \boldsymbol{\beta} \overset{d}{=} \prod_{i=1}^{n} z_i|y_i, \mathcal{M}, \boldsymbol{\beta},$$

factor into a set of n independent distributions with each depending on the data only through y_i. The distributions $z_i|y_i, \mathcal{M}, \boldsymbol{\beta}$ are obtained by reasoning as follows. Suppose that $y_i = 0$, then from Bayes theorem

$$\begin{aligned} f(z_i|y_i = 0, \mathcal{M}, \boldsymbol{\beta}) &\propto f_N(z_i|\boldsymbol{x}_i'\boldsymbol{\beta}, 1)\, f(y_i = 0|z_i, \mathcal{M}, \boldsymbol{\beta}) \\ &\propto f_N(z_i|\boldsymbol{x}_i'\boldsymbol{\beta}, 1)\, I[z_i \leqslant 0], \end{aligned}$$

because $f(y_i = 0|z_i, \mathcal{M}, \boldsymbol{\beta})$ is equal to one if z_i is negative and equal to zero otherwise, which is the definition of $I[z_i \leqslant 0]$. Hence, the information $y_i = 0$ simply serves to truncate the support of z_i. By a similar argument it is shown that the support of z_i is $(0, \infty)$ when conditioned on the event $y_i = 1$. Each of these truncated distributions is simulated by the formula given in Equation (5). This leads to the following algorithm.

Algorithm 8: Binary probit

(1) `Sample`

$$\boldsymbol{\beta} \sim \mathcal{N}_k\left(\boldsymbol{B}_n\left(\boldsymbol{B}_0^{-1}\boldsymbol{\beta}_0 + \sum_{i=1}^{n}\boldsymbol{x}_i z_i\right), \boldsymbol{B}_n = \left(\boldsymbol{B}_0^{-1} + \sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}\right)$$

(2) `Sample`

$$z_i \sim \begin{cases} \mathcal{TN}_{(-\infty, 0]}(\boldsymbol{x}_i'\boldsymbol{\beta}, 1) & \text{if } y_i = 0, \\ \mathcal{TN}_{(0, \infty)}(\boldsymbol{x}_i'\boldsymbol{\beta}, 1) & \text{if } y_i = 1, \end{cases} \quad i \leqslant n.$$

(3) `Goto 1.`

Albert and Chib (1993a) also extend this algorithm to the ordinal categorical data case where y_i can take one of the values $\{0, 1, \ldots, J\}$ according to the probabilities

$$\Pr(y_i \leqslant j|\boldsymbol{\beta}, \boldsymbol{\gamma}) = \Phi(\gamma_j - \boldsymbol{x}_i'\boldsymbol{\beta}), \quad j = 0, 1, \ldots, J. \tag{28}$$

In this model the $\{\gamma_j\}$ are category specific cut-points with γ_0 normalized to zero and γ_J to infinity. The remaining cut-points $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{J-1})$ are assumed to satisfy the order restriction $\gamma_1 \leqslant \cdots \leqslant \gamma_{J-1}$ which ensures that the cumulative probabilities are non-decreasing. For given data $y_1, \ldots, y_n$ from this model, the likelihood function is given by

$$f(\boldsymbol{y}|\mathcal{M}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \prod_{j=0}^{J} \prod_{i:y_i=j} \left[\Phi(\gamma_j - \boldsymbol{x}_i'\boldsymbol{\beta}) - \Phi(\gamma_{j-1} - \boldsymbol{x}_i'\boldsymbol{\beta})\right], \tag{29}$$

and the posterior density, under the prior $p(\boldsymbol{\beta}, \boldsymbol{\gamma})$, is proportional to $p(\boldsymbol{\beta}, \boldsymbol{\gamma})f(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\gamma})$. Posterior simulation is again feasible with the the introduction of latent variables $z_1, \ldots, z_n$, where $z_i|\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{x}_i\boldsymbol{\beta}, 1)$. A priori, we observe $y_i = j$ if the latent variable z_i falls in the interval $[\gamma_{j-1}, \gamma_j)$. Now the basic Albert and Chib MCMC scheme draws the latent data, regression parameters and cut-points in sequence. Given $y_i = j$, the sampling of the latent data z_i is from $\mathcal{TN}_{[\gamma_{j-1}, \gamma_j]}(\boldsymbol{x}_i'\boldsymbol{\beta}, 1)$ and the sampling of the parameters $\boldsymbol{\beta}$ is as in Algorithm 8. For the cut-points, Cowles (1996) and Nandram and Chen (1996) proposed that the cut-points be generated by the M–H algorithm, marginalized over $\boldsymbol{z}$. Subsequently, Albert and Chib (1998) simplified the latter step by transforming the cut-points $\boldsymbol{\gamma}$ so as to remove the ordering constraint. The transformation is defined by the one-to-one map

$$\delta_1 = \log \gamma_1;\ \delta_j = \log(\gamma_j - \gamma_{j-1}), \quad 2 \leqslant j \leqslant J - 1. \tag{30}$$

The advantage of working with $\boldsymbol{\delta}$ instead of $\boldsymbol{\gamma}$ is that the parameters of the tailored proposal density in the M–H step for $\boldsymbol{\delta}$ can be obtained by an unconstrained optimization and the prior $p(\boldsymbol{\delta})$ on $\boldsymbol{\delta}$ can be an unrestricted multivariate normal. The algorithm is defined as follows.

Algorithm 9: Ordinal probit

(1) M-H

(a) Calculate

$$\boldsymbol{m} = \arg\max_{\boldsymbol{\delta}} \log f(\boldsymbol{y}|\mathcal{M}, \boldsymbol{\beta}, \boldsymbol{\delta}),$$

and $\boldsymbol{V} = \{-\partial \log f(\boldsymbol{y}|\mathcal{M}, \boldsymbol{\beta}, \boldsymbol{\delta})/\partial\boldsymbol{\delta}\partial\boldsymbol{\delta}'\}^{-1}$, the negative inverse of the hessian at $\boldsymbol{m}$.

(b) Propose

$$\boldsymbol{\delta}' \sim f_T(\boldsymbol{\delta}|\boldsymbol{m}, \boldsymbol{V}, \xi).$$

(c) Calculate

$$\alpha = \min\left\{\frac{p(\boldsymbol{\delta}')f(\boldsymbol{y}|\mathcal{M}, \boldsymbol{\beta}, \boldsymbol{\delta}')}{p(\boldsymbol{\delta})f(\boldsymbol{y}|\mathcal{M}, \boldsymbol{\beta}, \boldsymbol{\delta})}\frac{f_T(\boldsymbol{\delta}|\boldsymbol{m}, \boldsymbol{V}, \xi)}{f_T(\boldsymbol{\delta}'|\boldsymbol{m}, \boldsymbol{V}, \xi)}, 1\right\}.$$

(d) Move to $\boldsymbol{\delta}'$ with probability α. Transform the new $\boldsymbol{\delta}$ to $\boldsymbol{\gamma}$ via the inverse map $\gamma_j = \sum_{i=1}^{j} \exp(\delta_i)$, $1 \leqslant j \leqslant J - 1$.

(2) Sample

$$z_i \sim \mathcal{TN}_{[\gamma_{j-1}, \gamma_j]}(\boldsymbol{x}_i'\boldsymbol{\beta}, 1) \quad \text{if} \quad y_i = j, \quad i \leqslant n.$$

(3) Sample

$$\boldsymbol{\beta} \sim \mathcal{N}_k\left(\boldsymbol{B}_n\left(\boldsymbol{B}_0^{-1}\boldsymbol{\beta}_0 + \sum_{i=1}^{n}\boldsymbol{x}_i z_i\right), \boldsymbol{B}_n = \left(\boldsymbol{B}_0^{-1} + \sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}\right).$$

(4) Goto 1.

8.5. Tobit censored regression

Consider now a model in the class of the Tobit family in which the data y_i is generated by

$$z_i|\mathcal{M}, \boldsymbol{\beta}, \sigma^2 \sim \mathcal{N}(\boldsymbol{x}_i'\boldsymbol{\beta}, \sigma^2).$$
$$y_i = \max(0, z_i), \quad 1 \leqslant i \leqslant n,$$

indicating that the observation z_i is observed only when z_i is positive. This model gives rise to a mixed discrete-continuous distribution with a point mass of $[1 - \Phi(\boldsymbol{x}_i'\boldsymbol{\beta}/\sigma)]$ at zero and a density $f_N(y_i|\boldsymbol{x}_i'\boldsymbol{\beta}, \sigma^2)$ on $(0, \infty)$. The likelihood function is given by

$$\prod_{i \in C}\{1 - \Phi(\boldsymbol{x}_i'\boldsymbol{\beta}/\sigma)\}\prod_{i \in C'}(\sigma^{-2})\exp\left\{-\frac{1}{2\sigma^2}(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2\right\},$$

where C is the set of censored observations and Φ is the c.d.f. of the standard normal random variable.

A MCMC procedure for this model is developed by Chib (1992) while Wei and Tanner (1990a) discuss a related approach for a model that arises in survival analysis. A set of tractable full conditional distributions is obtained by including the vector $\boldsymbol{z} = (z_i)$, $i \in C$ in the sampling. Let $\boldsymbol{y}_z = (y_{zi})$ be a $n \times 1$ vector with ith component y_i if the ith observation is not censored and z_i if it is censored. Now apply the Gibbs sampling algorithm with blocks $(\boldsymbol{\beta}, \sigma^2, \boldsymbol{z})$ and associated full conditional distributions

$$\boldsymbol{\beta}|\boldsymbol{y}, \mathcal{M}, \boldsymbol{z}, \sigma^2; \quad \sigma^2|\boldsymbol{y}, \mathcal{M}, \boldsymbol{z}, \boldsymbol{\beta} \quad \text{and} \quad \boldsymbol{z}|\boldsymbol{y}, \mathcal{M}, \boldsymbol{\beta}, \sigma^2.$$

The first two of these distributions follow from the results for linear regression with Gaussian errors (with y_{zi} used in place of y_i) and the third distribution, analogous to the probit case, is truncated normal on the interval $(-\infty, 0]$.

Algorithm 10: Tobit censored regression

```
(1) Sample
        $$\boldsymbol{\beta} \sim \mathcal{N}_k \left( \boldsymbol{B}_n \left( \boldsymbol{B}_0^{-1}\boldsymbol{\beta}_0 + \sigma^{-2} \sum_{t=1}^{n} \boldsymbol{x}_t' y_{zi} \right), \boldsymbol{B}_n = \left( \boldsymbol{B}_0^{-1} + \sigma^{-2} \sum_{t=1}^{n} \boldsymbol{x}_t \boldsymbol{x}_t' \right)^{-1} \right).$$
(2) Sample
        $$\sigma^2 \sim \mathcal{IG}\left\{ \frac{\upsilon_0 + n}{2}, \frac{\delta_0 + \sum_{i=1}^{n}(y_{zi} - \boldsymbol{x}_i'\boldsymbol{\beta})^2}{2} \right\}.$$
(3) Sample
        $$z_i \sim \mathcal{TN}_{(-\infty, 0]}(\boldsymbol{x}_i'\boldsymbol{\beta}, \sigma^2), \quad i \in C.$$
(4) Goto 1.
```

8.6. Regression with change point

Suppose that $\boldsymbol{y} = \{y_1, y_2, \ldots, y_n\}$ is a time series such that the density of y_t given $\boldsymbol{Y}_{t-1} = (y_1, \ldots, y_{t-1})$ is specified as

$$y_t|\mathcal{M}, \boldsymbol{Y}_{t-1}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma_1^2, \sigma_2^2, \tau \sim \begin{cases} \mathcal{N}(\boldsymbol{x}_t'\boldsymbol{\beta}_1, \sigma_1^2) & \text{if } t \leqslant \tau, \\ \mathcal{N}(\boldsymbol{x}_t'\boldsymbol{\beta}_2, \sigma_2^2) & \text{if } \tau < t, \end{cases}$$

where τ is an unknown change point. The objective is to estimate the parameter vectors $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$, the regression variances $\sigma^2 = (\sigma_1^2, \sigma_2^2)$ and the change point τ.

An analysis of such models from a MCMC perspective was initiated by Carlin, Gelfand and Smith (1992). It is based on the inclusion of the change point τ in the MCMC sampling. Stephens (1994) generalized the approach of Carlin, Gelfand and Smith for models with multiple change points by including each of the unobserved change points in the sampling. In this generalization, however, the step that involves the simulation of the change points conditioned on the parameters and the data can be computationally very demanding when the sample size n is large. A different approach to multiple change point problems which is computationally simpler is developed by Chib (1998). An important aspect of the MCMC approach for change point problems is that it can be easily adapted for binary and count data.

Assume that

$$\boldsymbol{\beta}_j \sim \mathcal{N}_k(\boldsymbol{\beta}_0, \boldsymbol{B}_0); \quad \sigma_j^2 \sim \mathcal{IG}\left(\frac{\nu_0}{2}, \frac{\delta_0}{2}\right); \quad \tau \sim \text{Unif}\{a_0, a_0+1, \ldots, b_0\},$$

where τ follows a discrete uniform distribution on the integers $\{a_0, b_0\}$. Then the posterior density is

$$\pi(\boldsymbol{\beta}, \sigma^2, \tau \mid \boldsymbol{y}, \mathcal{M}) \propto p(\boldsymbol{\beta}) p(\sigma^2) p(\tau) \prod_{t \leqslant \tau} \phi(y_t \mid \boldsymbol{x}_t' \boldsymbol{\beta}_1, \sigma_1^2) \prod_{\tau < t} \phi(y_t \mid \boldsymbol{x}_t' \boldsymbol{\beta}_2, \sigma_2^2).$$

Conditional on τ the data splits into two parts and the conditional distributions of the regression parameters are obtained from the regression updates of Algorithm 5. On the other hand, given the regression parameters, the full conditional distribution of τ is concentrated on $\{a_0, b_0\}$ with mass function

$$\Pr(\tau = k \mid \boldsymbol{y}, \mathcal{M}, \boldsymbol{\beta}, \sigma^2) \propto \prod_{t \leqslant k} \phi(y_t \mid \boldsymbol{x}_t' \boldsymbol{\beta}_1, \sigma_1^2) \prod_{k < t} \phi(y_t \mid \boldsymbol{x}_t' \boldsymbol{\beta}_2, \sigma_2^2).$$

The normalizing constant of this mass function is the sum of the right hand side over k.

Algorithm 11: Regression with change point

(1) Sample for $j = 1,2$

$$\begin{aligned}
&\boldsymbol{\beta}_j \sim \mathcal{N}_k(\hat{\boldsymbol{\beta}}_j, \boldsymbol{B}_j),\\
&\hat{\boldsymbol{\beta}}_j = \boldsymbol{B}_j \left(\boldsymbol{B}_0^{-1} \boldsymbol{\beta}_0 + \sigma_j^{-2} \sum_{t=l_j}^{u_j} \boldsymbol{x}_t' y_t \right),\\
&\boldsymbol{B}_j = \left(\boldsymbol{B}_0^{-1} + \sigma^{-2} \sum_{t=l_j}^{u_j} \boldsymbol{x}_t \boldsymbol{x}_t' \right)^{-1},\\
&l_j = 1 + (j-1)\tau;\ u_j = \tau + (j-1)(n-\tau).
\end{aligned}$$

(2) Sample for $j = 1,2$

$$\begin{aligned}
&\sigma_j^2 \sim \mathcal{IG}\left\{ \frac{\upsilon_0 + n_j}{2}, \frac{\delta_0 + \sum_{t=l_j}^{u_j} (y_t - \boldsymbol{x}_t' \boldsymbol{\beta}_j)^2}{2} \right\},\\
&n_j = \tau + (j-1)(n - 2\tau).
\end{aligned}$$

```
(3) Calculate for $k = a_0, a_0 + 1, \dots, b_0$
```

$$p_k \propto \prod_{t \leqslant k} \phi(y_t | \boldsymbol{x}_t' \boldsymbol{\beta}_1, \sigma_1^2) \prod_{k < t} \phi(y_t | \boldsymbol{x}_t' \boldsymbol{\beta}_2, \sigma_2^2).$$

```
(4) Sample
```

$$\tau \sim \{p_{a_0}, p_{a_0+1}, \dots, p_{b_0}\}.$$

```
(5) Goto 1.
```

8.7. Autoregressive time series

Consider the model

$$y_t = \boldsymbol{x}_t' \boldsymbol{\beta} + \epsilon_t, \quad 1 \leqslant t \leqslant n,$$

where the error is generated by the stationary AR(p) process

$$\epsilon_t - \phi_1 \epsilon_{t-1} - \cdots - \phi_p \epsilon_{t-p} = u_t \quad \text{or} \quad \phi(L)\, \epsilon_t = u_t,$$

where $u_t \sim$ i.i.d. $\mathcal{N}(0, \sigma^2)$ and $\phi(L) = 1 - \phi_1 L - \cdots - \phi_p L^p$ is a polynomial in the lag operator L. One interesting complication in this model is that the parameters $\boldsymbol{\phi} = (\phi_1, \dots, \phi_p)$, due to the stationarity assumption, are restricted to lie in the region $S_{\boldsymbol{\phi}}$ of $\Re^p$ where the roots of $\phi(L)$ are all outside the unit circle. Chib and Greenberg (1994), based on Chib (1993), derive a multiple-block Metropolis–Hastings MCMC algorithm for this model in which the proposal densities for $\boldsymbol{\beta}$ and σ^2 are the respective full conditional densities while that of $\boldsymbol{\phi}$ is a normal density constructed from the observations y_t, $t \geqslant p + 1$.

Denote the first p observations as $\boldsymbol{Y}_p = (y_1, \dots, y_p)'$ and $\boldsymbol{X}_p = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_p)'$ and let $y_t^* = \phi(L) y_t$ and $\boldsymbol{x}_t^* = \phi(L) \boldsymbol{x}_t$, $t \geqslant p + 1$. Also define the p dimensional matrix $\boldsymbol{\Sigma}_p$ through the matrix equation

$$\boldsymbol{\Sigma}_p = \boldsymbol{\Phi} \boldsymbol{\Sigma}_p \boldsymbol{\Phi}' + \boldsymbol{e}_1(p)\, \boldsymbol{e}_1(p)',$$

where

$$\boldsymbol{\Phi} = \begin{pmatrix} \boldsymbol{\phi}_{-p}' & \phi_p \\ \boldsymbol{I}_{p-1} & \boldsymbol{0} \end{pmatrix},$$

$\boldsymbol{e}_1(p) = (1, 0, \dots, 0)'$ and $\boldsymbol{\phi}_{-p} = (\phi_1, \dots, \phi_{p-1})'$. Let the cholesky factorization of $\boldsymbol{\Sigma}_p$ be $\boldsymbol{Q}\boldsymbol{Q}'$ and define $\boldsymbol{Y}_p^* = \boldsymbol{Q}^{-1} \boldsymbol{Y}_p$ and $\boldsymbol{X}_p^* = \boldsymbol{Q}^{-1} \boldsymbol{X}_p$ which are functions of $\boldsymbol{\phi}$. Finally define $e_t = y_t - \boldsymbol{x}_t' \boldsymbol{\beta}$, $t \geqslant p + 1$.

One can now proceed by noting that given $\boldsymbol{\phi}$, updates of $\boldsymbol{\beta}$ and σ^2 follow from the model

$$
\begin{aligned}
y_t^* | \mathcal{M}, \boldsymbol{\beta}, \sigma^2 &\sim \mathcal{N}(\boldsymbol{x}_t^{*\prime}\boldsymbol{\beta}, \sigma^2), \quad t \geqslant 1,\\
\boldsymbol{\beta} &\sim \mathcal{N}_k(\boldsymbol{\beta}_0, \boldsymbol{B}_0),\\
\sigma^2 &\sim \mathcal{IG}\left(\frac{\upsilon_0}{2}, \frac{\delta_0}{2}\right),
\end{aligned}
$$

while conditioned on $(\boldsymbol{\beta}, \sigma^2)$, and the assumption that the prior density of $\boldsymbol{\phi}$ is $\mathcal{N}(\boldsymbol{\phi}_0, G_0)$ truncated to the region S_ϕ, the full conditional of $\boldsymbol{\phi}$ is

$$
\pi(\boldsymbol{\phi} | \boldsymbol{y}, \mathcal{M}, \boldsymbol{\beta}, \sigma^2) \propto \Psi(\boldsymbol{\phi}) \times \mathcal{N}_p(\hat{\boldsymbol{\phi}}, \boldsymbol{V}) I_{S_\phi},
$$

where

$$
\Psi(\boldsymbol{\phi}) = |\boldsymbol{\Sigma}_p|^{-1/2} \exp\left\{ -\frac{1}{2\sigma^2} (\boldsymbol{Y}_p - \boldsymbol{X}_p\boldsymbol{\beta}) \boldsymbol{\Sigma}_p^{-1} (\boldsymbol{Y}_p - \boldsymbol{X}_p\boldsymbol{\beta}) \right\},
$$

$\hat{\boldsymbol{\phi}} = \boldsymbol{V}(\boldsymbol{G}_0^{-1}\boldsymbol{\phi}_0 + \sum_{t=p+1}^n \boldsymbol{E}_t\boldsymbol{e}_t)$, $\boldsymbol{V} = (\boldsymbol{G}_0^{-1} + \sigma^{-2}\sum_{t=p+1}^n \boldsymbol{E}_t\boldsymbol{E}_t^\prime)^{-1}$, $\boldsymbol{E}_t = (e_{t-1}, \ldots, e_{t-p})^\prime$. To sample this density the proposal density is specified as

$$
q(\boldsymbol{\phi} | \boldsymbol{y}, \mathcal{M}, \boldsymbol{\beta}, \sigma^2) = \mathcal{N}_p(\boldsymbol{\phi} | \hat{\boldsymbol{\phi}}, \boldsymbol{V}).
$$

With this tailored proposal density the probability of move just involves $\Psi(\boldsymbol{\phi})$, leading to a M–H step that is both fast (because it entails the calculation of a function based on the first p observations and not the entire sample) and highly efficient (because the proposal density is matched to the target).

Algorithm 12: Regression with autoregressive errors

```
(1) Calculate $(y_t^*, \boldsymbol{x}_t^*)$, $t \leqslant n$.
(2) Sample
        $\boldsymbol{\beta} \sim \mathcal{N}_k\left(\boldsymbol{B}_n(\boldsymbol{B}_0^{-1}\boldsymbol{\beta}_0 + \sigma^{-2}\sum_{t=1}^{n}\boldsymbol{x}_t^{*\prime}y_t^*), \boldsymbol{B}_n = (\boldsymbol{B}_0^{-1} + \sigma^{-2}\sum_{t=1}^{n}\boldsymbol{x}_t^*\boldsymbol{x}_t^{*\prime})^{-1}\right).$
(3) Sample
        $\sigma^2 \sim \mathcal{IG}\left\{\frac{\upsilon_0 + n}{2}, \frac{\delta_0 + \sum_{i=1}^{n}(y_t^* - \boldsymbol{x}_i^{*\prime}\boldsymbol{\beta})^2}{2}\right\}.$
(4) M-H
    (a) Calculate
            $\hat{\boldsymbol{\phi}} = \boldsymbol{V}(\boldsymbol{G}_0^{-1}\boldsymbol{\phi}_0 + \sigma^{-2}\sum_{t=p+1}^{n}\boldsymbol{E}_t^\prime e_t); \quad \boldsymbol{V} = (\boldsymbol{G}_0^{-1} + \sigma^{-2}\sum_{t=p+1}^{n}\boldsymbol{E}_t\boldsymbol{E}_t^\prime)^{-1}.$
    (b) Propose
            $\boldsymbol{\phi}' \sim \mathcal{N}_p(\hat{\boldsymbol{\phi}}, \boldsymbol{V})$.
    (c) Calculate
            $\alpha = \min\left\{1, \frac{\Psi(\boldsymbol{\phi}') I_{S_{\phi'}}}{\Psi(\boldsymbol{\phi})}\right\}$.
    (d) Move to $\boldsymbol{\phi}'$ with probability $\alpha$.
(5) Goto 1.
```

8.8. Hidden Markov models

In this subsection we consider the MCMC-based analysis of hidden Markov models (or Markov mixture models or Markov switching models). The general model is described as

$$\begin{aligned} y_t | \boldsymbol{Y}_{t-1}, \mathcal{M}, s_t = k, \boldsymbol{\theta} &\sim f(y_t | \boldsymbol{Y}_{t-1}, \mathcal{M}, \boldsymbol{\theta}_k), \quad k = 1, \ldots, m, \\ s_t | s_{t-1}, \boldsymbol{P} &\sim \text{Markov}(\boldsymbol{P}, \pi_1), \\ \boldsymbol{\theta} &\sim \pi, \\ \boldsymbol{p}_i &\sim \text{Dirichlet}(\alpha_{i1}, \ldots, \alpha_{im}), \quad i \leqslant m, \end{aligned}$$

where $s_t \in \{1, \ldots, m\}$ is an *unobservable* random variable which evolves according to a Markov process with transition matrix $\boldsymbol{P} = \{p_{ij}\}$, with $p_{ij} = \Pr(s_t = j | s_{t-1} = i)$, and initial distribution π_1 at $t = 1$, f is a density or mass function, $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m)$ are the parameters of f under each possible value of s_t, and $\boldsymbol{p}_i$ is the ith row of $\boldsymbol{P}$ that is assumed to have a Dirichlet prior distribution with parameters $(\alpha_{i1}, \ldots, \alpha_{im})$. For identifiability reasons, the Markov chain of s_t is assumed to be time-homogeneous, irreducible, and aperiodic.

The MCMC analysis of such models was initiated by Albert and Chib (1993b) in the context of a more general model than the one above where the conditional density of the data depends not just on s_t but also on the previous values $\{s_{t-1}, \ldots, s_{t-r}\}$, as in the model of Hamilton (1989). The approach relies on augmenting the parameter space to include the unobserved states and simulating $\pi(\boldsymbol{S}_n, \boldsymbol{\theta}, \boldsymbol{P} | \boldsymbol{y}, \mathcal{M})$ via the conditional distributions

$$s_t | \boldsymbol{y}, \mathcal{M}, \boldsymbol{S}_{(-t)}, \boldsymbol{\theta}, \boldsymbol{P} (t \leqslant n); \quad \boldsymbol{\theta} | \boldsymbol{y}, \mathcal{M}, \boldsymbol{S}_n, \boldsymbol{P}; \ \{\boldsymbol{p}_i\} | \boldsymbol{y}, \mathcal{M}, \boldsymbol{\theta}, \boldsymbol{S}_n,$$

where $\boldsymbol{S}_n = (s_1, \ldots, s_n)$ denotes the entire collection of states. Robert, Celeux and Diebolt (1993) and McCulloch and Tsay (1994) developed a similar approach for the simpler model in which only the current state s_t appears in the density of y_t while Billio, Monfort and Robert (1999) consider ARMA models with Markov switching.

Chib (1996), whose approach we now follow, modifies the first set of blocks of the above scheme to sample the states jointly from

$$\boldsymbol{S}_n|\boldsymbol{y}, \mathcal{M}, \boldsymbol{\theta}, \boldsymbol{P},$$

in one block. This leads to a more efficient MCMC algorithm. The sampling of $\boldsymbol{S}_n$ is achieved by *one* forward and backward pass through the data. In the forward pass, one recursively produces the sequence of mass functions $\{p(s_t|\boldsymbol{Y}_t, \mathcal{M}, \boldsymbol{\theta}, \boldsymbol{P})\}$ $(t \leqslant n)$ as follows: assume that the function $p(s_{t-1}|\boldsymbol{Y}_{t-1}, \mathcal{M}, \boldsymbol{\theta}, \boldsymbol{P})$ is available. Then, one obtains $p(s_t|\boldsymbol{Y}_t, \mathcal{M}, \boldsymbol{\theta}, \boldsymbol{P})$ by calculating

$$p(s_t|\boldsymbol{Y}_{t-1}, \mathcal{M}, \boldsymbol{\theta}, \boldsymbol{P}) = \sum_{l=1}^{m} p(s_t|\, s_{t-1} = l, \boldsymbol{\theta}, \boldsymbol{P}) \times p(s_{t-1} = l|\boldsymbol{Y}_{t-1}, \boldsymbol{\theta}, \boldsymbol{P}),$$

followed by

$$p(s_t|\boldsymbol{Y}_t, \mathcal{M}, \boldsymbol{\theta}, \boldsymbol{P}) = \frac{p(s_t|\boldsymbol{Y}_{t-1}, \mathcal{M}, \boldsymbol{\theta}, \boldsymbol{P}) \times f(y_t|\boldsymbol{Y}_{t-1}, \mathcal{M}, \boldsymbol{\theta}_{s_t}, \boldsymbol{P})}{\sum_{l=1}^{m} p(s_t = l|\boldsymbol{Y}_{t-1}, \mathcal{M}, \boldsymbol{\theta}, \boldsymbol{P}) \times f(y_t|\boldsymbol{Y}_{t-1}, \mathcal{M}, \boldsymbol{\theta}_l, \boldsymbol{P})}.$$

These forward recursions can be initialized at $t = 1$ by setting $p(s_1|\boldsymbol{Y}_0, \mathcal{M}, \boldsymbol{\theta}, \boldsymbol{P})$ to be the stationary distribution of the chain (the left eigenvector corresponding to the eigenvalue of one).

Then, in the backward pass one simulates $\boldsymbol{S}_n$ by the method of composition, first simulating s_n from $s_n|\boldsymbol{y}, \mathcal{M}, \boldsymbol{\theta}, \boldsymbol{P}$ and then the s_t's using the probability mass functions

$$p(s_t = k|\boldsymbol{y}, \mathcal{M}, \boldsymbol{S}^{t+1}, \boldsymbol{\theta}, \boldsymbol{P}) = \frac{p(s_t = k|\boldsymbol{Y}_t, \mathcal{M}, \boldsymbol{\theta}, \boldsymbol{P}) \times p(s_{t+1}|s_t = k, \boldsymbol{P})}{\sum_{l=1}^{m} p(s_t = l|\boldsymbol{Y}_t, \mathcal{M}, \boldsymbol{\theta}, \boldsymbol{P}) \times p(s_{t+1}|s_t = l, \boldsymbol{P})},$$
$$k \leqslant m, \quad t \leqslant n-1,$$

where $\boldsymbol{S}^{t+1} = (s_{t+1}, \ldots, s_n)$ consists of the simulated values from earlier steps and the second term of the numerator is the Markov transition probability, which is picked off from the column of $\boldsymbol{P}$ determined by the simulated value of s_{t+1}.

Given the simulated vector $\boldsymbol{S}_n$, the data separates into m non-contiguous pieces and the simulation of $\boldsymbol{\theta}_k$ is from the full conditional distribution

$$\pi(\boldsymbol{\theta}_k) \prod_{t:\, s_t = k} f(y_t|\boldsymbol{Y}_{t-1}, \mathcal{M}, \boldsymbol{\theta}, \boldsymbol{P}).$$

Depending on the form of f and p this may belong to a named distribution. Otherwise, this distribution is sampled by a M–H step. Finally, the last distribution depends simply on $\boldsymbol{S}_n$ with each row $\boldsymbol{p}_i$ of $\boldsymbol{P}$ independently an updated Dirichlet distribution:

$$\boldsymbol{p}_i|\boldsymbol{S}_n \sim \mathcal{D}(\alpha_{i1} + n_{i1}, \ldots, \alpha_{i1} + n_{im}), \quad (i \leqslant m),$$

where n_{ik} is the total number of *one-step* transitions from state i to state k in the vector $\boldsymbol{S}_n$.

Algorithm 13: Hidden Markov model

```
(1) Calculate and store for $t = 1, 2, \ldots, n$
        $p(s_t | \boldsymbol{Y}_t, \mathcal{M}, \boldsymbol{\theta}, \boldsymbol{P})$.
(2) Sample
        $s_n \sim p(s_n | \boldsymbol{y}, \mathcal{M}, \boldsymbol{\theta}, \boldsymbol{P})$.
(3) Sample for $t = n-1, n-2, \ldots, 1$
        $s_t \sim p(s_t | \boldsymbol{y}, \mathcal{M}, \boldsymbol{S}^{t+1}, \boldsymbol{\theta}, \boldsymbol{P})$.
(4) Sample for $k = 1, \ldots, m$
        $\boldsymbol{\theta}_k \propto \pi(\boldsymbol{\theta}_k) \prod_{t : s_t = k} f(y_t | \boldsymbol{Y}_{t-1}, \mathcal{M}, \boldsymbol{\theta}, \boldsymbol{P})$.
(5) Sample for $i = 1, 2, \ldots, m$
        $\boldsymbol{p}_i \sim \mathcal{D}(\alpha_{i1} + n_{i1}, \ldots, \alpha_{i1} + n_{im})$.
(6) Goto 1.
```

8.9. State space models

Consider next a linear state space model in which a scalar observation y_t is generated as

$$
\begin{aligned}
y_t | \mathcal{M}, \boldsymbol{\theta}_t &\sim \mathcal{N}(\boldsymbol{x}_t' \boldsymbol{\theta}_t, \sigma^2), \\
\boldsymbol{\theta}_t | \boldsymbol{\theta}_{t-1} &\sim \mathcal{N}_m(\boldsymbol{G} \boldsymbol{\theta}_{t-1}, \boldsymbol{\Psi}), \quad 1 \leqslant t \leqslant n, \\
\sigma^2 &\sim \mathcal{IG}\left(\frac{\upsilon_0}{2}, \frac{\delta_0}{2}\right), \\
\boldsymbol{\Psi}^{-1} &\sim \mathcal{W}_m(\rho_0, \boldsymbol{R}_0),
\end{aligned}
$$

where $\boldsymbol{\theta}_t$ is an $m \times 1$ state vector and $\boldsymbol{G}$ is assumed known. For nonlinear versions of this model, a MCMC fitting approach is provided by Carlin, Polson and Stoffer (1992). It is based on the inclusion of the variables $\{\boldsymbol{\theta}_t\}$ in the sampling followed by one-at-a-time sampling of $\boldsymbol{\theta}_t$ given $\boldsymbol{\theta}_{-t}$ (the remaining $\boldsymbol{\theta}_t$'s) and $(\sigma^2, \boldsymbol{\Psi})$. For the linear version presented above, Carter and Kohn (1994) and Fruhwirth-Schnatter (1994) show that a reduced blocking scheme involving the joint simulation of $\{\boldsymbol{\theta}_t\}$ is possible and desirable, because the $\boldsymbol{\theta}_t$'s are correlated by construction, while de Jong and Shephard (1995) provide an important alternative procedure called the simulation smoother that is particularly useful if $\boldsymbol{\Psi}$ is not positive definite or if the dimension m of the state vector is large. Carter and Kohn (1996) and Shephard (1994) also consider models, called conditionally Gaussian state space models, that have Gaussian observation densities conditioned on a discrete or continuous variable s_t. An example of this is provided below in Section 8.10. Chib and Greenberg (1995b) consider hierarchical and vector versions of the above model while additional issues related to the fitting and parameterization of state space models are considered by Pitt and Shephard (1997).

The MCMC implementation for this model is based on the distributions

$$\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_n | \boldsymbol{y}, \mathcal{M}, \sigma^2, \boldsymbol{\Psi}; \quad \sigma^2 | \boldsymbol{y}, \mathcal{M}, \{\boldsymbol{\theta}_t\}, \boldsymbol{\Psi}; \quad \boldsymbol{\Psi}^{-1} | \boldsymbol{y}, \mathcal{M}, \{\boldsymbol{\theta}_t\}.$$

To see how the $\boldsymbol{\theta}_t$'s are sampled, write the joint distribution as

$$\begin{aligned} & p(\boldsymbol{\theta}_n | \boldsymbol{y}, \mathcal{M}, \sigma^2, \boldsymbol{\Psi}) \times p(\boldsymbol{\theta}_{n-1} | \boldsymbol{y}, \mathcal{M}, \boldsymbol{\theta}_n, \sigma^2, \boldsymbol{\Psi}) \\ & \quad \times \cdots \times p(\boldsymbol{\theta}_1 | \boldsymbol{y}, \mathcal{M}, \boldsymbol{\theta}_2, \dots, \boldsymbol{\theta}_n, \sigma^2, \boldsymbol{\Psi}), \end{aligned}$$

where, on letting $\boldsymbol{\theta}^s = (\boldsymbol{\theta}_s, \dots, \boldsymbol{\theta}_n)$, $\boldsymbol{Y}_s = (y_1, \dots, y_s)$ and $\boldsymbol{Y}^s = (y_s, \dots, y_n)$ for $s \leqslant n$, the typical term is

$$p(\boldsymbol{\theta}_t | \boldsymbol{y}, \mathcal{M}, \boldsymbol{\theta}^{t+1}, \sigma^2, \boldsymbol{\Psi}) \propto p(\boldsymbol{\theta}_t | \boldsymbol{Y}_t, \mathcal{M}, \sigma^2, \boldsymbol{\Psi}) p(\boldsymbol{\theta}_{t+1} | \boldsymbol{\theta}_t, \mathcal{M}, \sigma^2, \boldsymbol{\Psi}),$$

due to the fact that $(\boldsymbol{Y}^{t+1}, \boldsymbol{\theta}^{t+1})$ is independent of $\boldsymbol{\theta}_t$ given $(\boldsymbol{\theta}_{t+1}, \sigma^2, \boldsymbol{\Psi})$. The first density on the right hand side is Gaussian with moments given by the Kalman filter recursions. The second density is Gaussian with moments $\boldsymbol{G\theta}_t$ and $\boldsymbol{\Psi}$. By completing the square in $\boldsymbol{\theta}_t$ the moments of $p(\boldsymbol{\theta}_t | \boldsymbol{y}, \mathcal{M}, \boldsymbol{\theta}^{t+1}, \sigma^2, \boldsymbol{\Psi})$ can be derived. Then, the joint distribution $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_n | \boldsymbol{y}, \mathcal{M}, \sigma^2, \boldsymbol{\Psi}$ can be sampled by the method of composition.

Algorithm 14: Gaussian state space

```
(1) Kalman filter
    (a) Calculate for $t = 1, 2, \dots, n$

            $\hat{\boldsymbol{\theta}}_{t|t-1} = \boldsymbol{G}\hat{\boldsymbol{\theta}}_{t-1|t-1}$,                    $\boldsymbol{R}_{t|t-1} = \boldsymbol{G}\boldsymbol{R}_{t-1|t-1}\boldsymbol{G}' + \boldsymbol{\Psi}$,
            $f_{t|t-1} = \boldsymbol{x}_t'\boldsymbol{R}_{t|t-1}\boldsymbol{x}_t + \sigma^2$,                $\boldsymbol{K}_t = \boldsymbol{R}_{t|t-1}\boldsymbol{x}_t f_{t|t-1}^{-1}$,
            $\hat{\boldsymbol{\theta}}_{t|t} = \hat{\boldsymbol{\theta}}_{t|t-1} + \boldsymbol{K}_t(y_t - \boldsymbol{x}_t'\hat{\boldsymbol{\theta}}_{t|t-1})$,   $\boldsymbol{R}_{t|t} = (\boldsymbol{I} - \boldsymbol{K}_t\boldsymbol{x}_t')\boldsymbol{R}_{t|t-1}$,
            $\boldsymbol{M}_t = \boldsymbol{R}_{t|t}\boldsymbol{G}'\boldsymbol{R}_{t+1|t}^{-1}$,                 $\boldsymbol{R}_t = \boldsymbol{R}_{t|t} - \boldsymbol{M}_t\boldsymbol{R}_{t+1|t}\boldsymbol{M}_t'$.

    (b) Store

            $\hat{\boldsymbol{\theta}}_{t|t}$; $\boldsymbol{M}_t$; $\boldsymbol{R}_t$.

(2) Simulation step
    (a) Sample

            $\boldsymbol{\theta}_n \sim \mathcal{N}_m(\hat{\boldsymbol{\theta}}_{n|n}, \boldsymbol{R}_{n|n})$.
```

(b) Sample for $t = n-1, n-2, \ldots, 1$

$$\boldsymbol{\theta}_t \sim \mathcal{N}_m(\hat{\boldsymbol{\theta}}_t, \boldsymbol{R}_t), \quad \hat{\boldsymbol{\theta}}_t = \hat{\boldsymbol{\theta}}_{t\,|\,t} + \boldsymbol{M}_t \left(\boldsymbol{\theta}_{t+1} - \boldsymbol{G}\hat{\boldsymbol{\theta}}_{t\,|\,t}\right).$$

(3) Sample

$$\sigma^2 \sim \mathcal{IG}\left\{\frac{\upsilon_0 + n}{2}, \frac{\delta_0 + \sum_{i=1}^{n}(y_t - \boldsymbol{x}_i'\boldsymbol{\theta}_t)^2}{2}\right\}.$$

(4) Sample

$$\boldsymbol{\Psi}^{-1} \sim \mathcal{W}_m\left[\rho_0 + n, \left\{\boldsymbol{R}_0^{-1} + \sum_{t=1}^{n}(\boldsymbol{\theta}_t - \boldsymbol{G}\boldsymbol{\theta}_{t-1})(\boldsymbol{\theta}_t - \boldsymbol{G}\boldsymbol{\theta}_{t-1})'\right\}^{-1}\right].$$

(5) Goto 1.

8.10. Stochastic volatility model

Suppose that time series observations $\{y_t\}$ are generated by the stochastic volatility (SV) model [see, for example, Taylor (1994), Shephard (1996), and Ghysels, Harvey and Renault (1996)]

$$y_t = \exp(h_t/2)u_t, \quad h_t = \mu + \phi(h_{t-1} - \mu) + \sigma\eta_t, \quad t \leqslant n,$$

where $\{h_t\}$ is the latent log-volatility of y_t and $\{u_t\}$ and $\{\eta_t\}$ are white noise standard normal random variables. This is an example of a state space model in which the state variable h_t appears non-linearly in the observation equation. The model can be extended to include covariates in the observation and evolution equations and to include a heavy-tailed, non-Gaussian distribution for u_t. The MCMC analysis of this model was initiated by Jacquier, Polson and Rossi (1994) based on the general approach of Carlin, Polson and Stoffer (1992). If we let $\boldsymbol{\theta} = (\phi, \mu, \sigma^2)$, then the algorithm of Jacquier, Polson and Rossi (1994) is based on the $(n+3)$ full conditional distributions

$$\begin{aligned}&h_t|\boldsymbol{y}, \mathcal{M}, h_{-t}, \boldsymbol{\theta}, \quad t = 1, 2, \ldots, n,\\ &\phi|\boldsymbol{y}, \mathcal{M}, \{h_t\}, \mu, \sigma^2; \quad \mu|\boldsymbol{y}, \mathcal{M}, \{h_t\}, \phi, \sigma^2; \; \sigma^2|\boldsymbol{y}, \mathcal{M}, \{h_t\}, \phi, \mu,\end{aligned}$$

where the latent variables h_t are sampled by a sequence of Metropolis–Hastings steps. Subsequently, Kim, Shephard and Chib (1998) discussed an alternative approach that leads to considerable improvements in the mixing of the Markov chain. The latter approach has been further refined by Chib, Nardari and Shephard (1998, 1999).

The idea behind the Kim, Shepard and Chib approach is to approximate the SV model by a conditionally Gaussian state space model with the introduction of

Table 1
Parameters of seven-component Gaussian mixture to approximate the distribution of $\log \chi^2_1$

s_t	q	m_{s_t}	$\upsilon^2_{s_t}$
1	0.00730	–11.40039	5.79596
2	0.10556	–5.24321	2.61369
3	0.00002	–9.83726	5.17950
4	0.04395	1.50746	0.16735
5	0.34001	–0.65098	0.64009
6	0.24566	0.52478	0.34023
7	0.25750	–2.35859	1.26261

multinomial random variables $\{s_t\}$ that follow a seven-point discrete distribution. Conditioned on $\{s_t\}$, the model is Gaussian and the variables h_t appear linearly in the observation equation. Then, the entire set of $\{h_t\}$ are sampled jointly conditioned on $\boldsymbol{\theta}$ and $\{s_t\}$ by either the simulation smoother of de Jong and Shephard (1995) or by the algorithm for simulating states given in Algorithm 14. Once the MCMC simulation is concluded the parameter draws are reweighted to correspond to the original non-linear model.

To begin with, reexpress the SV model as

$$y^*_t = h_t + z_t, \quad h_t = \mu + \phi(h_{t-1} - \mu) + \sigma\eta_t,$$

where $y^*_t = \ln(y^2_t)$ and $z_t = \log \varepsilon^2_t$ is distributed as a log of chi-squared random variable with one degrees of freedom. Now approximate the distribution of $y^*_t | h_t$ by a mixture of normal distributions. A very accurate representation is given by the mixture distribution

$$y^*_t | h_t, s_t \sim \mathcal{N}(m_{s_t} + h_t, \upsilon^2_{s_t}), \quad \Pr(s_t = i) = q_i, \quad i \leqslant 7, \quad t \leqslant n,$$

where $s_t \in (1, 2, \ldots, 7)$ is an unobserved component indicator with probability mass function $q = \{q_i\}$ and the parameters $\{q, m_{s_t}, \upsilon^2_{s_t}\}$ are as reported in Table 1. Now the parameters and the latent variables can be simulated by a *two block* MCMC algorithm defined by the distributions

$$(\boldsymbol{\theta}, h_1, \ldots, h_n) | \{y^*_t\}, \{s_t\},$$
$$\{s_t\} | \{y^*_t\}, \{h_t\}, \boldsymbol{\theta}.$$

where the first block is sampled by the method of composition by first drawing $\boldsymbol{\theta}$ from $\pi(\boldsymbol{\theta} | \{y^*_t\}, \{s_t\})$ by a M–H step followed by a draw of $\{h_t\}$ by the simulation smoother. In the former step the target distribution is

$$\begin{aligned}\pi(\boldsymbol{\theta} | \{y^*_t\}, \{s_t\}) &\propto p(\boldsymbol{\theta}) f(y^*_1, \ldots, y^*_n | \{s_t\}, \boldsymbol{\theta}) \\ &= p(\boldsymbol{\theta}) \prod_{t=1}^{n} f(y^*_t | \mathcal{F}^*_{t-1}, \{s_t\}, \boldsymbol{\theta}),\end{aligned}$$

where each one-step ahead density $f(y_t^*|\mathcal{F}_{t-1}^*, \{s_t\}, \boldsymbol{\theta})$ can be derived from the output of the Kalman filter recursions, adapted to the differing components, as indicated by the component vector $\{s_t\}$, and $p(\boldsymbol{\theta})$ is the prior density. For ϕ the prior can be taken to be the scaled beta density

$$p(\phi) = c\,(0.5(1+\phi))^{\phi^{(1)}-1}\,(0.5(1-\phi))^{\phi^{(2)}-1}, \quad \phi^{(1)}, \phi^{(2)} > 0.5, \tag{31}$$

where

$$c = 0.5\frac{\Gamma(\phi^{(1)}+\phi^{(2)})}{\Gamma(\phi^{(1)})\Gamma(\phi^{(2)})},$$

with prior mean of $2\phi^{(1)}/(\phi^{(1)}+\phi^{(2)}-1)$, while those on μ and σ^2 can be normal and inverse gamma densities, respectively.

Algorithm 15: Stochastic volatility

```
(1) Initialize $\{s_t\}$
(2) M-H
    (a) Calculate $\boldsymbol{m} = \arg\max_{\boldsymbol{\theta}} l(\boldsymbol{\theta})$ where

            $l(\boldsymbol{\theta}) = -\frac{1}{2}\sum_{t=1}^{n} \ln f_{t|t-1} - \frac{1}{2}\sum_{t=1}^{n} \frac{(y_t^* - m_{s_t} - \hat{h}_{t|t-1})^2}{f_{t|t-1}}$

        and $\boldsymbol{V} = \{-\partial^2 l(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'\}^{-1}$, the negative inverse of the
        hessian at $\boldsymbol{m}$, where $f_{t|t-1}$ and $\hat{h}_{t|t-1}$ are computed from the
        Kalman filter recursions

            $\hat{h}_{t|t-1} = \mu + \phi(\hat{h}_{t-1|t-1} - \mu)$,        $\boldsymbol{R}_{t|t-1} = \phi^2 \boldsymbol{R}_{t-1|t-1} + \sigma^2$,
            $f_{t|t-1} = \boldsymbol{R}_{t|t-1} + v_{s_t}^2$,                $\boldsymbol{K}_t = \boldsymbol{R}_{t|t-1} f_{t|t-1}^{-1}$,
            $\hat{h}_{t|t} = \hat{h}_{t|t-1} + \boldsymbol{K}_t(y_t^* - m_{s_t} - \hat{h}_{t|t-1})$, $\boldsymbol{R}_{t|t} = (1 - \boldsymbol{K}_t)\boldsymbol{R}_{t|t-1}$.

    (b) Propose

            $\boldsymbol{\theta}' \sim f_T(\boldsymbol{\theta}|\boldsymbol{m}, \boldsymbol{V}, \xi)$.

    (c) Calculate

            $\alpha = \min\left\{\frac{p(\boldsymbol{\theta}')\, l(\boldsymbol{\theta}')}{p(\boldsymbol{\theta})\, l(\boldsymbol{\theta})} \frac{f_T(\boldsymbol{\theta}|\boldsymbol{m}, \boldsymbol{V}, \xi)}{f_T(\boldsymbol{\theta}'|\boldsymbol{m}, \boldsymbol{V}, \xi)}, 1\right\}$.

    (d) Move to $\boldsymbol{\theta}'$ with probability $\alpha$.
(3) Sample $\{h_t\}$ using algorithm 13, or the simulation smoother
    algorithm, modified to include the components of the
    mixture selected by $\{s_t\}$.
```

(4) `Sample`

$$s_t \sim \Pr(s_t | y_t^*, h_t, \boldsymbol{\psi}) \propto \Pr(s_t) f_N(y_t^* | \mu_{s_t} + h_t, \upsilon_{s_t}^2).$$

(5) `Goto 2.`

8.11. Gaussian panel data models

For continuous clustered or panel data a common model formulation is that of Laird and Ware (1982)

$$\boldsymbol{y}_i | \mathcal{M}, \boldsymbol{\beta}, \boldsymbol{b}_i, \sigma^2 \sim \mathcal{N}_{n_i}(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{W}_i\boldsymbol{b}_i, \sigma^2\boldsymbol{I}_{n_i}), \quad \boldsymbol{b}_i | \boldsymbol{D} \sim \mathcal{N}_q(\boldsymbol{0}, \boldsymbol{D}),$$
$$\boldsymbol{D}^{-1} \sim \mathcal{W}_p(\rho_0, \boldsymbol{R}_0), \quad \boldsymbol{\beta} \sim \mathcal{N}_k(\boldsymbol{\beta}_0, \boldsymbol{B}_0), \quad \sigma^2 \sim \mathcal{IG}\left(\frac{\nu_0}{2}, \frac{\delta_0}{2}\right),$$

where $\boldsymbol{y}_i$ is a n_i vector of observations and the matrix $\boldsymbol{W}_i$ is a subset of $\boldsymbol{X}_i$. If $\boldsymbol{W}_i$ is a vector of units, then the model reduces to a panel model with intercept heterogeneity. If $\boldsymbol{W}_i = \boldsymbol{X}_i$, then the model becomes the random coefficient panel model.

Zeger and Karim (1991) and Wakefield et al. (1994) propose a Gibbs MCMC approach for this model that is based on including $\{\boldsymbol{b}_i\}$ in the sampling in conjunction with full blocking. This blocking scheme is not very desirable because the random effects and the fixed effects $\boldsymbol{\beta}$ tend to be highly correlated and treating them as separate blocks creates problems with mixing Gelfand, Sahu and Carlin (1995). To deal with this problem, Chib and Carlin (1999) suggest a number of reduced blocking schemes. One of the simplest proceeds by sampling $\boldsymbol{\beta}$ and $\{\boldsymbol{b}_i\}$ in one block by the method of composition: first sampling $\boldsymbol{\beta}$ marginalized over $\{\boldsymbol{b}_i\}$ and then sampling $\{\boldsymbol{b}_i\}$ conditioned on $\boldsymbol{\beta}$. What makes reduced blocking possible is the fact that the distribution of $\boldsymbol{y}_i$ marginalized over $\boldsymbol{b}_i$ is also Gaussian:

$$\boldsymbol{y}_i | \mathcal{M}, \boldsymbol{\beta}, \boldsymbol{D}, \sigma^2 \sim \mathcal{N}_{n_i}(\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{V}_i), \quad \boldsymbol{V}_i = \sigma^2\boldsymbol{I}_{n_i} + \boldsymbol{W}_i\boldsymbol{D}\boldsymbol{W}_i'.$$

The updated distribution of $\boldsymbol{\beta}$, marginalized over $\{\boldsymbol{b}_i\}$ is, therefore, easy to derive. The rest of the algorithm follows the steps of Wakefield et al. (1994). In particular, the sampling of the random effects is from independent normal distributions that are derived by treating $(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})$ as the "data," $\boldsymbol{b}_i$ as the regression coefficient and $\boldsymbol{b}_i \sim \mathcal{N}_q(\boldsymbol{0}, \boldsymbol{D})$ as the prior. The sampling of $\boldsymbol{D}^{-1}$ is from an Wishart distribution and that of σ^2 from an inverse gamma distribution.

Algorithm 16: Gaussian Panel

(1) `Sample`

$$\boldsymbol{\beta} \sim \mathcal{N}_k\left(\boldsymbol{B}_n(\boldsymbol{B}_0^{-1}\boldsymbol{\beta}_0 + \sum_{i=1}^{n} \boldsymbol{X}_i\boldsymbol{V}_i^{-1}\boldsymbol{y}_i), \boldsymbol{B}_n = (\boldsymbol{B}_0^{-1} + \sum_{i=1}^{n} \boldsymbol{X}_i\boldsymbol{V}_i^{-1}\boldsymbol{X}_i)^{-1}\right).$$

(2) `Sample`

$$\boldsymbol{b}_i \sim \mathcal{N}_q\left(\boldsymbol{D}_i \boldsymbol{W}_i' \sigma^{-2}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}), \boldsymbol{D}_i = (\boldsymbol{D} + \sigma^{-2}\boldsymbol{W}_i'\boldsymbol{W}_i)^{-1}\right), \quad i \leqslant n.$$

(3) `Sample`

$$\boldsymbol{D}^{-1} \sim \mathcal{W}_p\left\{\rho_0 + n, \left(\boldsymbol{R}_0^{-1} + \sum_{i=1}^{n} \boldsymbol{b}_i\boldsymbol{b}_i'\right)^{-1}\right\}.$$

(4) `Sample`

$$\sigma^2 \sim \mathcal{IG}\left(\frac{\nu_0 + \sum n_i}{2}, \frac{\delta_0 + \sum_{i=1}^{n} \| \boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{W}_i\boldsymbol{b}_i \|^2}{2}\right).$$

(5) `Goto 1.`

8.12. Multivariate binary data models

To model correlated binary data a canonical model is the multivariate probit (MVP). Let y_{ij} denote the binary response on the ith observation unit and jth variable, and let $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iJ})'$, $1 \leqslant i \leqslant n$, denote the collection of responses on all J variables. Then, under the MVP model the marginal probability of $y_{ij} = 1$ is

$$\Pr(y_{ij} = 1|\mathcal{M}, \boldsymbol{\beta}) = \Phi(\boldsymbol{x}_{ij}'\boldsymbol{\beta}_j),$$

and the joint probability that $\boldsymbol{Y}_i = \boldsymbol{y}_i$ conditioned on the parameters $(\boldsymbol{\beta}, \boldsymbol{\Sigma})$ is

$$\Pr(\boldsymbol{Y}_i = \boldsymbol{y}_i|\mathcal{M}, \boldsymbol{\beta}, \boldsymbol{\Sigma}) \equiv \Pr(\boldsymbol{y}_i|\mathcal{M}, \boldsymbol{\beta}, \boldsymbol{\Sigma}) = \int_{A_{iJ}} \cdots \int_{A_{i1}} \phi_J(\boldsymbol{t}|\boldsymbol{0}, \boldsymbol{\Sigma})\,\mathrm{d}t,$$

where as in the SUR model, $\boldsymbol{\beta}' = (\boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_J') \in \Re^k$, $k = \sum k_j$, but unlike the SUR model, the $J-$ matrix $\boldsymbol{\Sigma} = \{\sigma_{jk}\}$ is in correlation form (with units on the diagonal), and A_{ij} is the interval

$$A_{ij} = \begin{cases} (-\infty, \boldsymbol{x}_{ij}'\boldsymbol{\beta}_j) & \text{if } y_{ij} = 1, \\ \left[\boldsymbol{x}_{ij}'\boldsymbol{\beta}_j, \infty\right) & \text{if } y_{ij} = 0. \end{cases}$$

To simplify the MCMC implementation for this model Chib and Greenberg (1998) follow the general approach of Albert and Chib (1993a) and employ latent variables. Let

$$\boldsymbol{z}_i \sim \mathcal{N}_J(\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}),$$

with the observed data given by the sign of z_{ij}:

$$y_{ij} = I(z_{ij} > 0), \quad j = 1, \ldots, J,$$

where $I(A)$ is the indicator function of the event A. If we let $\boldsymbol{\sigma} = (\sigma_{21}, \sigma_{31}, \sigma_{32}, \ldots, \sigma_{JJ})$ denote the $J(J-1)/2$ distinct elements of $\boldsymbol{\Sigma}$, and let $\boldsymbol{z} = (z_1, \ldots, z_n)$ denote the latent values corresponding to the observed data $\boldsymbol{Y} = \{\boldsymbol{y}_i\}_{i=1}^{n}$, then the algorithm proceeds with the sampling of the augmented posterior density

$$\begin{aligned}\pi(\boldsymbol{\beta}, \boldsymbol{\sigma}, \boldsymbol{z}|\,\boldsymbol{y}, \mathcal{M}) &\propto p(\boldsymbol{\beta})p(\boldsymbol{\sigma})f(\boldsymbol{z}|\mathcal{M}, \boldsymbol{\beta}, \boldsymbol{\Sigma})\Pr(\boldsymbol{y}|\boldsymbol{z}, \boldsymbol{\beta}, \boldsymbol{\Sigma})\\ &\propto p(\boldsymbol{\beta})p(\boldsymbol{\sigma})\prod_{i=1}^{n}\{\phi_J(\boldsymbol{z}_i|\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma})\Pr(\boldsymbol{y}_i|\boldsymbol{z}_i, \boldsymbol{\beta}, \boldsymbol{\Sigma})\}, \boldsymbol{\beta} \in \Re^k, \boldsymbol{\sigma} \in C,\end{aligned}$$

where

$$\Pr(\boldsymbol{y}_i|\boldsymbol{z}_i, \boldsymbol{\beta}, \boldsymbol{\Sigma}) = \prod_{j=1}^{J}\{I(z_{ij} > 0)I(y_{ij} = 1) + I(z_{ij} \leqslant 0)I(y_{ij} = 0)\},$$

$p(\boldsymbol{\sigma})$ is a normal density truncated to the region C, and C is the set of values of $\boldsymbol{\sigma}$ that produce a positive definite correlation matrix $\boldsymbol{\Sigma}$.

Conditioned on $\{\boldsymbol{z}_i\}$ and $\boldsymbol{\Sigma}$, the update for $\boldsymbol{\beta}$ is as in the SUR model, while conditioned on $(\boldsymbol{\beta}, \boldsymbol{\Sigma})$, z_{ij} can be sampled one at a time conditioned on the other latent values from truncated normal distributions, where the region of truncation is either $(0, \infty)$ or $(-\infty, 0)$ depending on whether the corresponding y_{ij} is one or zero. The key step in the algorithm is the sampling of $\boldsymbol{\sigma}$, the unrestricted elements of $\boldsymbol{\Sigma}$, from the full conditional density $\pi(\boldsymbol{\sigma}|\mathcal{M}, \boldsymbol{z}, \boldsymbol{\beta}) \propto p(\boldsymbol{\sigma})\prod_{i=1}^{n}\phi_J(\boldsymbol{z}_i|\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma})$. This density, which is truncated to the complicated region C, is sampled by a M–H step with tailored proposal density $q(\boldsymbol{\sigma}|\mathcal{M}, \boldsymbol{z}, \boldsymbol{\beta}) = f_T(\boldsymbol{\sigma}|\boldsymbol{m}, \boldsymbol{V}, \xi)$ where

$$\begin{aligned}\boldsymbol{m} &= \arg\max_{\boldsymbol{\sigma} \in C}\sum_{i=1}^{n}\ln\phi_J(\boldsymbol{z}_i|\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}),\\ \boldsymbol{V} &= -\left\{\frac{\partial^2\sum_{i=1}^{n}\ln\phi_J(\boldsymbol{z}_i|\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma})}{\partial\boldsymbol{\sigma}\partial\boldsymbol{\sigma}'}\right\}_{\boldsymbol{\sigma}=\boldsymbol{m}}^{-1},\end{aligned}$$

are the mode and curvature of the target distribution, given the current values of the conditioning variables. Note that, as in Algorithm 12, no truncation is enforced on the proposal density.

Algorithm 17: Multivariate probit

```
(1) Sample for $i \leqslant n, j \leqslant J$
        $z_{ij} \sim \begin{cases} \mathcal{TN}_{(0,\infty)}(\mu_{ij}, v_{ij}) & \text{if } y_{ij} = 1, \\ \mathcal{TN}_{(-\infty,0])}(\mu_{ij}, v_{ij}) & \text{if } y_{ij} = 0, \end{cases}$
        $\mu_{ij} = \mathrm{E}(z_{ij}|Z_{i(-j)}, \boldsymbol{\beta}, \boldsymbol{\Sigma})$,
        $v_{ij} = \mathrm{Var}(z_{ij}|Z_{i(-j)}, \boldsymbol{\beta}, \boldsymbol{\Sigma})$.
```

```
(2) Sample
        $\boldsymbol{\beta} \sim \mathcal{N}_k \left( \boldsymbol{B}_n \left( \boldsymbol{B}_0^{-1}\boldsymbol{\beta}_0 + \sum_{i=1}^{n} \boldsymbol{X}_i' \boldsymbol{\Sigma}^{-1} \boldsymbol{z}_i \right), \boldsymbol{B}_n = \left( \boldsymbol{B}_0^{-1} + \sum_{i=1}^{n} \boldsymbol{X}_i' \boldsymbol{\Sigma}^{-1} \boldsymbol{X}_i^{-1} \right)^{-1} \right).$
(3) M-H
    (a) Calculate the parameters $(\boldsymbol{m}, \boldsymbol{V})$.
    (b) Propose
            $\boldsymbol{\sigma}' \sim f_T(\boldsymbol{\sigma} | \boldsymbol{m}, \boldsymbol{V}, \xi).$
    (c) Calculate
            $\alpha = \min \left\{ \frac{p(\boldsymbol{\sigma}') \prod_{i=1}^{n} \phi_J(z_i | \boldsymbol{X}_i \boldsymbol{\beta}, \boldsymbol{\Sigma}') I[\boldsymbol{\sigma}' \in C]}{p(\boldsymbol{\sigma}) \prod_{i=1}^{n} \phi_J(z_i | \boldsymbol{X}_i \boldsymbol{\beta}, \boldsymbol{\Sigma})} \frac{f_T(\boldsymbol{\sigma} | \boldsymbol{m}, \boldsymbol{V}, \xi)}{f_T(\boldsymbol{\sigma}' | \boldsymbol{m}, \boldsymbol{V}, \xi)}, 1 \right\}.$
    (d) Move to $\boldsymbol{\sigma}'$ with probability $\alpha$.
(4) Goto 1.
```

As an application of this algorithm consider a data set in which the multivariate binary responses are generated by a panel strucure. The data is concerned with the health effects of pollution on 537 children in Stuebenville, Ohio, each observed at ages 7, 8, 9 and 10 years, and the response variable is an indicator of wheezing status [Diggle, Liang and Zeger (1995)]. Suppose that the marginal probability of wheeze status of the ith child at the jth time point is specified as

$$\Pr(y_{ij} = 1 | \boldsymbol{\beta}) = \Phi(\beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \beta_3 x_{3ij}), \quad i \leqslant 537, j \leqslant 4,$$

where $\boldsymbol{\beta}$ is constant across categories, x_1 is the age of the child centered at nine years, x_2 is a binary indicator variable representing the mother's smoking habit during the first year of the study, and $x_3 = x_1 x_2$. Suppose that the Gaussian prior on $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4)$ is centered at zero with a variance of $10\boldsymbol{I}_k$ and let $p(\boldsymbol{\sigma})$ be the density of a normal distribution, with mean zero and variance $\boldsymbol{I}_6$, *restricted* to region that leads to a positive-definite correlation matrix, where $(\sigma_{21}, \sigma_{31}, \sigma_{32}, \sigma_{41}, \sigma_{42}, \sigma_{43})$. From 10 000 cycles of Algorithm 17 one obtains the following covariate effects and posterior distributions of the correlations.

Notice that the summary tabular output in Table 2 contains not only the posterior means and standard deviations of the parameters but also the 95% credibility intervals, all computed from the sampled draws. It may be seen from Figure 6 that the posterior distributions of the correlations are similar suggesting that an equicorrelated correlation structure might be appropriate for these data. This issue is considered more formally in Section 10.2 below.

Table 2
Covariate effects in the Ohio wheeze data: MVP model with unrestricted correlations[1]

β	Prior		Posterior[2]				
	Mean	Std. dev.	Mean	NSE	Std. dev.	Lower	Upper
$\boldsymbol{\beta}_1$	0.000	3.162	−1.108	0.001	0.062	−1.231	−0.985
$\boldsymbol{\beta}_2$	0.000	3.162	−0.077	0.001	0.030	−0.136	−0.017
$\boldsymbol{\beta}_3$	0.000	3.162	0.155	0.002	0.101	−0.043	0.352
$\boldsymbol{\beta}_4$	0.000	3.162	0.036	0.001	0.049	−0.058	0.131

[1] The results are based on 10 000 draws from Algorithm 17.

[2] NSE denotes the numerical standard error, lower is the 2.5th percentile and upper is the 97.5th percentile of the simulated draws.

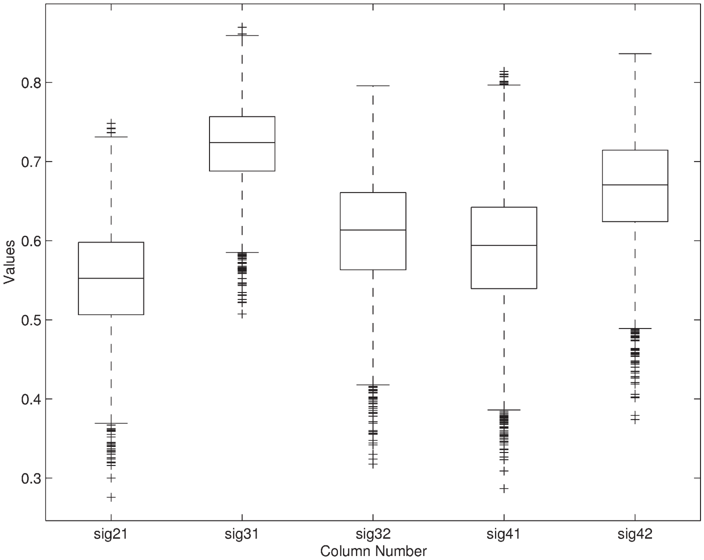

Fig. 6. Posterior boxplots of the correlations in the Ohio wheeze data: MVP model.

9. Sampling the predictive density

A fundamental goal of any statistical analysis is to predict a set of future or unobserved observations $\boldsymbol{y}_f$ given the current data $\boldsymbol{y}$ and the assumed model $\mathcal{M}$. In the Bayesian context this problem is solved by the calculation of the Bayesian prediction density which is defined as the distribution of $\boldsymbol{y}_f$ conditioned on $(\boldsymbol{y}, \mathcal{M})$ but marginalized over the parameters $\boldsymbol{\theta}$. More formally, the predictive density is defined as

$$f(\boldsymbol{y}_f|\boldsymbol{y}, \mathcal{M}) = \int f(\boldsymbol{y}_f|\boldsymbol{y}, \mathcal{M}, \boldsymbol{\theta})\, \pi(\boldsymbol{\theta}|\boldsymbol{y}, \mathcal{M})\, \mathrm{d}\boldsymbol{\theta}, \tag{32}$$

where $f(\boldsymbol{y}_f|\boldsymbol{y}, \mathcal{M}, \boldsymbol{\theta})$ is the conditional density of $\boldsymbol{y}_f$ given $(\boldsymbol{y}, \mathcal{M}, \boldsymbol{\theta})$ and the marginalization is with respect to the posterior density $\pi(\boldsymbol{\theta}|\boldsymbol{y}, \mathcal{M})$ of $\boldsymbol{\theta}$. In general, the predictive density is not available in closed form. However, in the context of MCMC problems that deliver a sample of (correlated) draws

$$\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(M)} \sim \pi(\boldsymbol{\theta}|\boldsymbol{y}, \mathcal{M}),$$

this is hardly a problem. One can utilize the posterior draws in conjunction with the method of composition to produce a sample of draws from the predictive density. This is done by appending a step at the end of the MCMC iterations where for each value $\boldsymbol{\theta}^{(j)}$ one simulates

$$\boldsymbol{y}_f^{(j)} \sim f(\boldsymbol{y}_f|\boldsymbol{y}, \mathcal{M}, \boldsymbol{\theta}^{(j)}), \quad j \leqslant M, \tag{33}$$

from the density of the observations, conditioned on $\boldsymbol{\theta}^{(j)}$. The collection of simulated values $\{\boldsymbol{y}_f^{(1)}, \ldots, \boldsymbol{y}_f^{(M)}\}$ is a sample from the Bayes prediction density $f(\boldsymbol{y}_f|\boldsymbol{y}, \mathcal{M})$. The simulated sample can be summarized in the usual way by the computation of sample averages and quantiles. Thus, to sample the prediction density one simply has to simulate the data generating process for each simulated value of the parameters.

In some problems, that have a latent data structure, a modified procedure to sample the predictive density may be necessary. Suppose that $\boldsymbol{z}_f$ denotes the latent data in the prediction period and $\boldsymbol{z}$ denote the latent data in the sample period. Let $\boldsymbol{\psi} = (\boldsymbol{\theta}, \boldsymbol{z})$ and suppose that the MCMC sampler produces the draws

$$\boldsymbol{\psi}^{(1)}, \ldots, \boldsymbol{\psi}^{(M)} \sim \pi(\boldsymbol{\psi}|\boldsymbol{y}, \mathcal{M}).$$

In this situation, the predictive density can be expressed as

$$f(\boldsymbol{y}_f|\boldsymbol{y}, \mathcal{M}) = \int f(\boldsymbol{y}_f|\boldsymbol{y}, \mathcal{M}, \boldsymbol{z}_f, \boldsymbol{\psi})\, \pi(\boldsymbol{z}_f|\boldsymbol{y}, \mathcal{M}, \boldsymbol{\psi})\, \pi(\boldsymbol{\psi}|\boldsymbol{y}, \mathcal{M})\, \mathrm{d}\boldsymbol{z}_f\, \mathrm{d}\boldsymbol{\psi}, \tag{34}$$

which may again be sampled by the method of composition where for each value $\boldsymbol{\psi}^{(j)} \sim \pi(\boldsymbol{\psi}|\boldsymbol{y}, \mathcal{M})$ one simulates

$$\boldsymbol{z}_f^{(j)} \sim \pi(\boldsymbol{z}_f|\boldsymbol{y}, \mathcal{M}, \boldsymbol{\psi}^{(j)}), \quad \boldsymbol{y}_f^{(j)} \sim f(\boldsymbol{y}_f|\boldsymbol{y}, \mathcal{M}, \boldsymbol{z}_f^{(j)}, \boldsymbol{\psi}^{(j)}).$$

The simulated values of $\boldsymbol{y}_f$ from this two step process are again from the predictive density.

To illustrate the one step procedure, suppose that one is interested in predicting $\boldsymbol{y}_f = (y_{n+1}, y_{n+2})$ from a regression model with autoregressive errors of order two where

$$y_t | \boldsymbol{Y}_{t-1}, \mathcal{M}, \boldsymbol{\beta}, \phi, \sigma^2 \sim \mathcal{N}(\phi_1 y_{t-1} + \phi_2 y_{t-2} + (\boldsymbol{x}_t - \phi_1 \boldsymbol{x}_{t-1} - \phi_2 \boldsymbol{x}_{t-2})' \boldsymbol{\beta}, \sigma^2).$$

Then, for each draw $(\boldsymbol{\beta}^{(j)}, \phi^{(j)}, \sigma^{2(j)})$ from Algorithm 12, one simulates $\boldsymbol{y}_f$ by sampling

$$y_{n+1}^{(j)} \sim \mathcal{N}(\phi_1^{(j)} y_n + \phi_2^{(j)} y_{n-1} + (\boldsymbol{x}_{n+1} - \phi_1^{(j)} \boldsymbol{x}_n - \phi_2^{(j)} \boldsymbol{x}_{n-1})' \boldsymbol{\beta}^{(j)}, \sigma^{2(j)})$$

and

$$y_{n+2}^{(j)} \sim \mathcal{N}(\phi_1^{(j)} y_{n+1}^{(j)} + \phi_2^{(j)} y_n + (\boldsymbol{x}_{n+2} - \phi_1^{(j)} \boldsymbol{x}_{n+1} - \phi_2^{(j)} \boldsymbol{x}_n)' \boldsymbol{\beta}^{(j)}, \sigma^{2(j)}).$$

The sample of simulated values $\{y_{n+1}^{(j)}, y_{n+2}^{(j)}\}$ from repeating this process is a sample from the (joint) predictive density.

As an example of the two step procedure consider a specific hidden Markov model in which

$$y_t | \boldsymbol{Y}_{t-1}, \mathcal{M}, \beta_0, \gamma, \sigma^2 \sim \mathcal{N}(\beta_0 + \gamma s_t, \sigma^2),$$

where $s_t \in \{0, 1\}$ is a unobserved state variable that follows a two-state Markov process with unknown transition probabilities

$$\boldsymbol{P} = \begin{pmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{pmatrix}.$$

In this case, suppose that Algorithm 13 has been used to deliver draws on $\boldsymbol{\psi} = (\beta_0, \gamma, \sigma^2, a, b, S_n)$. As described by Albert and Chib (1993b), to predict y_{n+1} we take each draw of $\boldsymbol{\psi}^{(j)}$ and sample

$$s_{n+1}^{(j)} \sim p(s_{n+1} | s_n^{(j)}, p_{11}^{(j)}, p_{22}^{(j)})$$

from the Markov chain (this is just a two point discrete distribution), and then sample

$$y_{n+1}^{(j)} \sim \mathcal{N}(\beta_0^{(j)} + \gamma^{(j)} s_{n+1}^{(j)}, \sigma^{2(j)}).$$

The next value $y_{n+2}^{(j)}$ is drawn in the same way after $s_{n+2}^{(j)}$ is simulated from the Markov chain $p(s_{n+2} | s_{n+1}^{(j)}, p_{11}^{(j)}, p_{22}^{(j)})$. These two steps can be iterated for any number of periods into the future and the whole process repeated for each simulated value of $\boldsymbol{\psi}$.

10. MCMC methods in model choice problems

10.1. Background

Consider the situation in which there are K possible models $\mathcal{M}_1, \ldots, \mathcal{M}_K$ for the observed data defined by the sampling densities $\{f(\boldsymbol{y}|\boldsymbol{\theta}_k, \mathcal{M}_k)\}$ and proper prior densities $\{p(\boldsymbol{\theta}_k|\mathcal{M}_k)\}$ and the objective is to find the evidence in the data for the different models. In the Bayesian approach this question is answered by placing prior probabilities $\Pr(\mathcal{M}_k)$ on each of the K models and using the Bayes calculus to find the posterior probabilities $\{\Pr(\mathcal{M}_1|\boldsymbol{y}), \ldots, \Pr(\mathcal{M}_K|\boldsymbol{y})\}$ conditioned on the data but marginalized over the unknowns $\boldsymbol{\theta}_k$. Specifically, the posterior probability of $\mathcal{M}_k$ is given by the expression

$$\begin{aligned}\Pr(\mathcal{M}_k|\boldsymbol{y}) &= \frac{\Pr(\mathcal{M}_k)\, m(\boldsymbol{y}|\mathcal{M}_k)}{\sum_{l=1}^{K} \Pr(\mathcal{M}_l)\, m(\boldsymbol{y}|\mathcal{M}_l)} \\ &\propto \Pr(\mathcal{M}_k)\, m(\boldsymbol{y}|\mathcal{M}_k), \quad (k \leqslant K),\end{aligned}$$

where

$$m(\boldsymbol{y}|\mathcal{M}_k) = \int f(\boldsymbol{y}|\boldsymbol{\theta}_k, \mathcal{M}_k) p(\boldsymbol{\theta}_k|\mathcal{M}_k)\, \mathrm{d}\boldsymbol{\theta}_k, \tag{35}$$

is the marginal density of the data and is called the marginal likelihood of $\mathcal{M}_k$. In words, the posterior probability of $\mathcal{M}_k$ is proportional to the prior probability of $\mathcal{M}_k$ times the marginal likelihood of $\mathcal{M}_k$. The evidence provided by the data about the models under consideration is summarized by the posterior probability of each model.

Often the posterior probabilities are summarized in terms of the posterior odds

$$\frac{\Pr(\mathcal{M}_i|\boldsymbol{y})}{\Pr(\mathcal{M}_j|\boldsymbol{y})} = \frac{\Pr(\mathcal{M}_i)}{\Pr(\mathcal{M}_j)} \frac{m(\boldsymbol{y}|\mathcal{M}_i)}{m(\boldsymbol{y}|\mathcal{M}_j)},$$

which provides the relative support for the two models. The ratio of marginal likelihoods in this expression is the Bayes factor of $\mathcal{M}_i$ vs $\mathcal{M}_j$.

If interest centers on the prediction of observables then it is possible to mix over the alternative predictive densities by utilizing the posterior probabilities as weights. More formally, the prediction density of a set of observations $\boldsymbol{y}_f$ marginalized over both $\{\boldsymbol{\theta}_k\}$ and $\{\mathcal{M}_k\}$ is given by

$$f(\boldsymbol{y}_f|\boldsymbol{y}) = \sum_{j=1}^{K} \Pr(\mathcal{M}_k|\boldsymbol{y}) f(\boldsymbol{y}_f|\boldsymbol{y}, \mathcal{M}_k),$$

where $f(\boldsymbol{y}_f|\boldsymbol{y}, \mathcal{M}_k)$ is the prediction density in Equation (34).

10.2. Marginal likelihood computation

A central problem in estimating the marginal likelihood is that it is an integral of the sampling density over the prior distribution of $\boldsymbol{\theta}_k$. Thus, MCMC methods, which deliver sample values from the posterior density, cannot be used to directly average the sampling density because that estimate would converge to

$$\int f(\boldsymbol{y}|\boldsymbol{\theta}_k, \mathcal{M}_k)\, p(\boldsymbol{\theta}_k|\boldsymbol{y}, \mathcal{M}_k)\, \mathrm{d}\boldsymbol{\theta}_k,$$

which is not the marginal likelihood. In addition, taking draws from the prior density to do the averaging produces an estimate that is simulation-consistent but highly inefficient because draws from the prior density are not likely to be in high density regions of the sampling density $f(\boldsymbol{y}|\boldsymbol{\theta}_k, \mathcal{M}_k)$. A natural way to correct this problem is by the method of importance sampling. If we let $h(\boldsymbol{\theta}_k|\mathcal{M}_k)$ denote a suitable importance sampling function, then the marginal likelihood can be estimated as

$$\begin{aligned} \hat{m}_I(\boldsymbol{y}|\mathcal{M}_k) &= M^{-1} \sum_{j=1}^{M} \frac{f(\boldsymbol{y}|\boldsymbol{\theta}_k^{(j)}, \mathcal{M}_k)\, p(\boldsymbol{\theta}_k^{(j)}|\mathcal{M}_k)}{h(\boldsymbol{\theta}_k^{(j)}|\mathcal{M}_k)}, \\ \boldsymbol{\theta}_k^{(j)} &\sim h(\boldsymbol{\theta}_k^{(j)}|\mathcal{M}_k) \quad (j \leqslant M). \end{aligned}$$

This method is useful when it can be shown that the ratio is bounded, which can be difficult to check in practice, and when the sampling density is not expensive to compute which, unfortunately, is often not true. We mention that if the importance sampling function is taken to be the unnormalized posterior density then that leads to

$$\hat{m}_{\mathrm{NR}} = \left[\frac{1}{M} \sum_{j=1}^{M} \left\{ \frac{1}{f(\boldsymbol{y}|\boldsymbol{\theta}_k^{(j)}, \mathcal{M}_k)\, p(\boldsymbol{\theta}_k^{(j)}|\mathcal{M}_k)} \right\} \right]^{-1},$$

the harmonic mean of the likelihood values. This estimate, proposed by Newton and Raftery (1994), can be unstable because the inverse likelihood does not have finite variance. Gelfand and Dey (1994) propose a modified stable estimator

$$\hat{m}_{\mathrm{GD}} = \left[\frac{1}{M} \sum_{j=1}^{M} \left\{ \frac{h(\boldsymbol{\theta}^{(j)})}{f(\boldsymbol{y}|\boldsymbol{\theta}_k^{(j)}, \mathcal{M}_k)\, p(\boldsymbol{\theta}_k^{(j)}|\mathcal{M}_k)} \right\} \right]^{-1},$$

where $h(\boldsymbol{\theta})$ is a density with tails thinner than the product of the prior and the likelihood. Unfortunately, this estimator is difficult to apply in models with latent or missing data.

The Laplace method for integrals can be used to provide a non-simulation based estimate of the marginal likelihood. Let d_k denote the dimension of $\boldsymbol{\theta}_k$ and let $\hat{\boldsymbol{\theta}}_k$

denote the posterior mode of $\boldsymbol{\theta}_k$, and $\boldsymbol{\Sigma}_k$ the inverse of the negative Hessian of $\ln\{f(\boldsymbol{y}|\boldsymbol{\theta}_k, \mathcal{M}_k)p(\boldsymbol{\theta}_k|\mathcal{M}_k)\}$ evaluated at $\hat{\boldsymbol{\theta}}_k$. Then the Laplace estimate of marginal likelihood, on the customary log base ten scale, is given by

$$\log \hat{m}_L(\boldsymbol{y}|\mathcal{M}_k) = (d_k/2)\log(2\pi) + (1/2)\log\det(\boldsymbol{\Sigma}_k) + \log f(\boldsymbol{y}|\hat{\boldsymbol{\theta}}_k, \mathcal{M}_k) + \log p(\hat{\boldsymbol{\theta}}_k|\mathcal{M}_k).$$

The Laplace estimate has a large sample justification and can be shown to equal the true value upto an error that goes to zero in probability at the rate n^{-1}.

Both the importance method and the Laplace estimate may be considered as the traditional methods for computing the marginal likelihood. More recent methods exploit two additional facts about the marginal likelihood. The first that the marginal likelihood is the normalizing constant of the posterior density and therefore under this view the Bayes factor can be interpreted as the ratio of two normalizing constants. There is a large literature in physics (in a quite different context, however) on precisely the latter problem stemming from Bennett (1976). This literature was adapted in the mid 1990's for statistical problems by Meng and Wong (1996) utilizing the bridge sampling method and by Chen and Shao (1997) based on umbrella sampling. The techniques presented in these papers, although based on the work in physics, contain modifications of the ideas to handle problems such as models with differing dimensions. DiCiccio, Kass, Raftery and Wasserman (1997) present a comparative analysis of the bridge sampling method in relation to other competing methods of computing the marginal likelihood. At this time, however, the bridge sampling method and its refinements have not found significant use in applications perhaps because the methods are quite involved and because simpler methods are available.

Another approach that deals with the estimation of Bayes factors, again in the context of nested models, is due to Verdinelli and Wasserman (1995) and is called the Savage–Dickey density ratio method. Suppose a model is defined by a parameter $\boldsymbol{\theta} = (\boldsymbol{\omega}, \boldsymbol{\psi})$ and the first model $\mathcal{M}_1$ is defined by the restriction $\boldsymbol{\omega} = \boldsymbol{\omega}_0$ and the second model $\mathcal{M}_2$ by letting $\boldsymbol{\omega}$ be unrestricted. Then, it can be shown that the Bayes factor is given by

$$B_{12} = \frac{\pi(\boldsymbol{\omega}_0|\boldsymbol{y}, \mathcal{M}_2)}{\pi(\boldsymbol{\omega}_0|\mathcal{M}_2)} E\left\{\frac{p(\boldsymbol{\psi}|\mathcal{M}_1)}{p(\boldsymbol{\psi}|\mathcal{M}_1, \boldsymbol{\omega}_0)}\right\},$$

where the expectation is with respect to $\pi(\boldsymbol{\psi}|\boldsymbol{y}, \mathcal{M}_2, \boldsymbol{\omega}_0)$. If $\pi(\boldsymbol{\omega}_0|\boldsymbol{y}, \mathcal{M}_2, \boldsymbol{\psi})$ is available in closed form then $\pi(\boldsymbol{\omega}_0|\boldsymbol{y}, \mathcal{M}_2)$ can be estimated by the Rao–Blackwell method and the second expectation by taking draws from the posterior $\pi(\boldsymbol{\psi}|\boldsymbol{y}, \mathcal{M}_2, \boldsymbol{\omega}_0)$, which can be obtained by the method of reduced runs discussed below, and averaging the ratio of prior densities. This method provides a simple approach for nested models but the method is not efficient if the dimensions of the two models are substantially different because then the ordinate $\pi(\boldsymbol{\omega}_0|\boldsymbol{y}, \mathcal{M}_2)$ tends to be small and the simulated values used to average the ratio tend to be in low density regions.

The second fact about marginal likelihoods, highlighted in a paper by Chib (1995), is that the marginal likelihood by virtue of being the normalizing constant of the posterior density can be expressed as

$$m(\boldsymbol{y}|\mathcal{M}_k) = \frac{f(\boldsymbol{y}|\boldsymbol{\theta}_k, \mathcal{M}_k)p(\boldsymbol{\theta}_k|\mathcal{M}_k)}{\pi(\boldsymbol{\theta}_k|\boldsymbol{y}, \mathcal{M}_k)}. \tag{36}$$

This expression is an identity in $\boldsymbol{\theta}_k$ because the left hand side is free of $\boldsymbol{\theta}_k$. Chib (1995) refers to it as the basic marginal likelihood identity (BMI). Based on this expression an estimate of the marginal likelihood on the log-scale is given by

$$\log \hat{m}(\boldsymbol{y}|\mathcal{M}_k) = \log f(\boldsymbol{y}|\boldsymbol{\theta}_k^*, \mathcal{M}_k) + \log p(\boldsymbol{\theta}_k^*|\mathcal{M}_k) - \log \hat{\pi}(\boldsymbol{\theta}_k^*|\boldsymbol{y}, \mathcal{M}_k), \tag{37}$$

where $\boldsymbol{\theta}_k^*$ denotes an arbitrarily chosen point and $\hat{\pi}(\boldsymbol{\theta}_k^*|\boldsymbol{y}, \mathcal{M}_k)$ is the estimate of the posterior density at that single point. Two points should be noted. First, this estimate requires only one evaluation of the likelihood function. This is particularly useful in situations where repeated evaluation of the likelihood function is computationally expensive. Second, to increase the computational efficiency, the point $\boldsymbol{\theta}_k^*$ should be taken to be a high density point under the posterior.

To estimate the posterior ordinate one utilizes the MCMC output in conjunction with a marginal/conditional decomposition. To simplify notation, drop the model subscript k and suppose that the parameter vector is blocked into B blocks as $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_B$. In addition, let $\boldsymbol{z}$ denote additional variables (latent or missing data) that may be included in the simulation to clarify the structure of the full conditional distributions. Also let $\boldsymbol{\psi}_i = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_i)$ and $\boldsymbol{\psi}^i = (\boldsymbol{\theta}_i, \ldots, \boldsymbol{\theta}_B)$ denote the list of blocks upto i and the set of blocks from i to B, respectively. Now write the posterior ordinate at the point $\boldsymbol{\theta}^*$ by the law of total probability as

$$\pi(\boldsymbol{\theta}^*|\boldsymbol{y}, \mathcal{M}) = \pi(\boldsymbol{\theta}_1^*|\boldsymbol{y}, \mathcal{M}) \times \cdots \times \pi(\boldsymbol{\theta}_i^*|\boldsymbol{y}, \mathcal{M}, \boldsymbol{\psi}_{i-1}^*) \times \cdots \times \pi(\boldsymbol{\theta}_B^*|\boldsymbol{y}, \mathcal{M}, \boldsymbol{\psi}_{B-1}^*), \tag{38}$$

where the first term in this expression is the marginal density of $\boldsymbol{\theta}_1$ evaluated at $\boldsymbol{\theta}_1^*$, and the typical term is of the form

$$\pi(\boldsymbol{\theta}_i^*|\boldsymbol{y}, \mathcal{M}, \boldsymbol{\psi}_{i-1}^*) = \int \pi(\boldsymbol{\theta}_i^*|\boldsymbol{y}, \mathcal{M}, \boldsymbol{\psi}_{i-1}^*, \boldsymbol{\psi}^{i+1}, \boldsymbol{z})\, \pi(\boldsymbol{\psi}^{i+1}, \boldsymbol{z}|\boldsymbol{y}, \mathcal{M}, \boldsymbol{\psi}_{i-1}^*)\, \mathrm{d}\boldsymbol{\psi}^{i+1}\, \mathrm{d}\boldsymbol{z}.$$

This may be called a *reduced conditional ordinate.* It is important to bear in mind that in finding the reduced conditional ordinate one must integrate only over $(\boldsymbol{\psi}^{i+1}, \boldsymbol{z})$ and that the integrating measure is conditioned on $\boldsymbol{\psi}_{i-1}^*$.

Assume that the normalizing constants of each full conditional density is known, an assumption that is relaxed below. Then, the first term of Equation (38) can be estimated by the Rao–Blackwell method. To estimate the typical reduced conditional ordinate, Chib (1995) defines a *reduced MCMC* run consisting of the full conditional distributions

$$\begin{aligned}\{&\pi(\boldsymbol{\theta}_i|\boldsymbol{y},\mathcal{M},\boldsymbol{\psi}^*_{i-1},\boldsymbol{\psi}^{i+1},\boldsymbol{z});\ \cdots;\ \pi(\boldsymbol{\theta}_B|\boldsymbol{y},\mathcal{M},\boldsymbol{\psi}^*_{i-1},\boldsymbol{\theta}_i,\ldots,\boldsymbol{\theta}_{B-1},\boldsymbol{z});\\ &\pi(\boldsymbol{z}|\boldsymbol{y},\mathcal{M},\boldsymbol{\psi}^*_{i-1},\boldsymbol{\psi}^i)\},\end{aligned}\tag{39}$$

where the blocks in $\boldsymbol{\psi}_{i-1}$ are set equal to $\boldsymbol{\psi}^*_{i-1}$. By MCMC theory, the draws on $(\boldsymbol{\psi}^{i+1},\boldsymbol{z})$ from this run are from the distribution $\pi(\boldsymbol{\psi}^{i+1},\boldsymbol{z}|\boldsymbol{y},\mathcal{M},\boldsymbol{\psi}^*_{i-1})$ and so the reduced conditional ordinate can be estimated as the average

$$\hat{\pi}(\boldsymbol{\theta}^*_i|\boldsymbol{y},\mathcal{M},\boldsymbol{\psi}^*_{i-1}) = M^{-1}\sum_{j=1}^{M}\pi(\boldsymbol{\theta}^*_i|\boldsymbol{y},\mathcal{M},\boldsymbol{\psi}^*_{i-1},\boldsymbol{\psi}^{i+1,(j)},\boldsymbol{z}^{(j)}),$$

over the simulated values of $\boldsymbol{\psi}^{i+1}$ and $\boldsymbol{z}$ from the reduced run. Each subsequent reduced conditional ordinate that appears in the decomposition (38) can be estimated in the same way though, conveniently, with fewer and fewer distributions appearing in the reduced runs. Given the marginal and reduced conditional ordinates, the Chib estimate of the marginal likelihood on the log scale is defined as

$$\log\hat{m}(\boldsymbol{y}|\mathcal{M}) = \log f(\boldsymbol{y}|\boldsymbol{\theta}^*,\mathcal{M}) + \log p(\boldsymbol{\theta}^*) - \sum_{i=1}^{B}\log\hat{\pi}(\boldsymbol{\theta}^*_i|\boldsymbol{y},\mathcal{M},\boldsymbol{\psi}^*_{i-1}),\tag{40}$$

where $f(\boldsymbol{y}|\boldsymbol{\theta}^*,\mathcal{M})$ is the density of the data marginalized over the latent data $\boldsymbol{z}$.

It is worth noting that an alternative approach to estimate the posterior ordinate is developed by Ritter and Tanner (1992) in the context of Gibbs MCMC chains with fully known full conditional distributions. If one lets

$$K_G(\boldsymbol{\theta},\boldsymbol{\theta}^*|\boldsymbol{y},\mathcal{M}) = \prod_{i=1}^{B}\pi(\boldsymbol{\theta}^*_k|\boldsymbol{y},\mathcal{M},\boldsymbol{\theta}^*_1,\ldots,\boldsymbol{\theta}^*_{k-1},\boldsymbol{\theta}_{k+1},\ldots,\boldsymbol{\theta}_B),$$

denote the Gibbs transition kernel, then by virtue of the fact that the Gibbs chain satisfies the invariance condition $\pi(\boldsymbol{\theta}^*|\boldsymbol{y},\mathcal{M}) = \int K_G(\boldsymbol{\theta},\boldsymbol{\theta}^*|\boldsymbol{y},\mathcal{M})\,\pi(\boldsymbol{\theta}|\boldsymbol{y},\mathcal{M})\,\mathrm{d}\boldsymbol{\theta}$, one can obtain the posterior ordinate by averaging the transition kernel over draws from the posterior distribution:

$$\hat{\pi}(\boldsymbol{\theta}^*|\boldsymbol{y},\mathcal{M}) = M^{-1}\sum_{g=1}^{M}K_G(\boldsymbol{\theta}^{(g)},\boldsymbol{\theta}^*|\boldsymbol{y},\mathcal{M}).$$

This estimate only requires draws from the full Gibbs run but when $\boldsymbol{\theta}$ is high dimensional and the model contains latent variables, this estimate is less accurate than Chib's posterior density decomposition method.

It should be observed that the above methods of estimating the posterior ordinate require knowledge of the normalizing constants of each full conditional density. What can be done when this condition does not hold? DiCiccio, Kass, Raftery and Wasserman (1997) and Chib and Greenberg (1998) suggest the use of kernel smoothing in this case. Suppose, for example, that the problem occurs in the distribution of the ith block. Then, the draws on $\boldsymbol{\theta}_i$ from the reduced MCMC run in Equation (39) can be smoothed by kernel methods to find the ordinate at $\boldsymbol{\theta}_i^*$. This approach should only be used when the dimension of the recalcitrant block is not large. A more general technique has recently been developed by Chib and Jeliazkov (2001). The first main result of the paper is that if sampling is done in one block by the M–H algorithm then the posterior ordinate can be written as

$$\pi(\boldsymbol{\theta}^*|\boldsymbol{y},\mathcal{M}) = \frac{E_1\left\{\alpha(\boldsymbol{\theta},\boldsymbol{\theta}^*|\boldsymbol{y},\mathcal{M})\,q(\boldsymbol{\theta},\boldsymbol{\theta}^*|\boldsymbol{y},\mathcal{M})\right\}}{E_2\left\{\alpha(\boldsymbol{\theta}^*,\boldsymbol{\theta}|\boldsymbol{y},\mathcal{M})\right\}},$$

where the numerator expectation E_1 is with respect to the distribution $\pi(\boldsymbol{\theta}|\boldsymbol{y},\mathcal{M})$ and the denominator expectation E_2 is with respect to the proposal density of $\boldsymbol{\theta}$ conditioned on $\boldsymbol{\theta}^*$, $q(\boldsymbol{\theta}^*,\boldsymbol{\theta}|\boldsymbol{y},\mathcal{M})$, and $\alpha(\boldsymbol{\theta},\boldsymbol{\theta}^*|\boldsymbol{y},\mathcal{M})$ is the probability of move in the M–H step. This expression implies that a simulation consistent estimate of the posterior ordinate can be defined as

$$\hat{\pi}(\boldsymbol{\theta}^*|\boldsymbol{y}) = \frac{M^{-1}\sum_{g=1}^{M}\alpha(\boldsymbol{\theta}^{(g)},\boldsymbol{\theta}^*|\boldsymbol{y},\mathcal{M})\,q(\boldsymbol{\theta}^{(g)},\boldsymbol{\theta}^*|\boldsymbol{y},\mathcal{M})}{J^{-1}\sum_{j=1}^{M}\alpha(\boldsymbol{\theta}^*,\boldsymbol{\theta}^{(j)}|\boldsymbol{y},\mathcal{M})}, \tag{41}$$

where $\{\boldsymbol{\theta}^{(g)}\}$ are the given draws from the posterior distribution while the draws $\boldsymbol{\theta}^{(j)}$ in the denominator are from $q(\boldsymbol{\theta}^*,\boldsymbol{\theta}|\boldsymbol{y},\mathcal{M})$, given the fixed value $\boldsymbol{\theta}^*$. The second main result of the paper is that in the context of the multiple block M–H algorithm the reduced conditional ordinate can be expressed as

$$\begin{aligned}&\pi(\boldsymbol{\theta}_i^*|\boldsymbol{y},\mathcal{M},\boldsymbol{\theta}_1^*,\ldots,\boldsymbol{\theta}_{i-1}^*)\\&\quad= \frac{E_1\left\{\alpha(\boldsymbol{\theta}_i,\boldsymbol{\theta}_i^*|\boldsymbol{y},\mathcal{M},\boldsymbol{\psi}_{i-1}^*,\boldsymbol{\psi}^{i+1})q_i(\boldsymbol{\theta}_i,\boldsymbol{\theta}_i^*|\boldsymbol{y},\mathcal{M},\boldsymbol{\psi}_{i-1}^*,\boldsymbol{\psi}^{i+1})\right\}}{E_2\left\{\alpha(\boldsymbol{\theta}_i^*,\boldsymbol{\theta}_i|\boldsymbol{y},\mathcal{M},\boldsymbol{\psi}_{i-1}^*,\boldsymbol{\psi}^{i+1})\right\}},\end{aligned} \tag{42}$$

where E_1 is the expectation with respect to $\pi(\boldsymbol{\theta}_i,\boldsymbol{\psi}^{i+1}|\boldsymbol{y},\mathcal{M},\boldsymbol{\psi}_{i-1}^*)$ and E_2 that with respect to the product measure $\pi(\boldsymbol{\psi}^{i+1}|\boldsymbol{y},\mathcal{M},\boldsymbol{\psi}_i^*)\,q_i(\boldsymbol{\theta}_i^*,\boldsymbol{\theta}_i|\boldsymbol{y},\mathcal{M},\boldsymbol{\psi}_{i-1}^*,\boldsymbol{\psi}^{i+1})$. The quantity $\alpha(\boldsymbol{\theta}_i,\boldsymbol{\theta}_i^*|\boldsymbol{y},\mathcal{M},\boldsymbol{\psi}_{i-1}^*,\boldsymbol{\psi}^{i+1})$ is the usual *conditional* M–H probability of move. The two expectations can be estimated from the output of the reduced runs in an obvious way. An example of this technique in action is provided next.

Consider the data set that was introduced in Section 8 in connection with the multivariate probit model. In this setting, the full conditonal density of the correlations is not in tractable form. Assume as before that the marginal probability of wheeze is given by

$$\Pr(y_{ij}=1|\mathcal{M}_k,\boldsymbol{\beta}) = \Phi(\beta_0+\beta_1x_{1ij}+\beta_2x_{2ij}+\beta_3x_{3ij}), \quad i\leqslant 537,\ j\leqslant 4,$$

where, as before, the dependence of $\boldsymbol{\beta}$ on the model is suppressed for convenience, x_1 is the age of the child centered at nine years, x_2 is a binary indicator variable representing the mother's smoking habit during the first year of the study, and $x_3 = x_1 x_2$. Now suppose that interest centers on three alternative models generated by three alternative correlation matrices. Let these models be defined as

- $\mathcal{M}_1$: Unrestricted $\boldsymbol{\Sigma}$ except for the unit constraints on the diagonal. In this case $\boldsymbol{\sigma}$ consists of six unknown elements.
- $\mathcal{M}_2$: Equicorrelated $\boldsymbol{\Sigma}$ where the correlations are all equal and described by a single parameter ρ.
- $\mathcal{M}_3$: Toeplitz $\boldsymbol{\Sigma}$ wherein the correlations depend on a single parameter ω but under the restriction that $\mathrm{Corr}(Z_{ik}, Z_{il}) = \omega^{|k-l|}$.

Assume that the prior on $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)$ is independent Gaussian with a mean of zero and a variance of ten. Also let the prior on the correlations $\boldsymbol{\sigma}$ be normal with mean of zero and covariance equal to the identity matrix (truncated to the region C) and that on ρ and ω be normal truncated to the interval $(-1, 1)$.

For each model, 10 000 iterations of Algorithm 17 are used to obtain the posterior sample and the posterior ordinate, using $\mathcal{M}_1$ for illustration, is computed as

$$\pi(\boldsymbol{\sigma}^*, \boldsymbol{\beta}^* | \boldsymbol{y}, \mathcal{M}_1) = \pi(\boldsymbol{\sigma}^* | \boldsymbol{y}, \mathcal{M}_1)\, \pi(\boldsymbol{\beta}^* | \boldsymbol{y}, \mathcal{M}_1, \boldsymbol{\Sigma}^*).$$

To estimate the marginal ordinate one can apply Equation (42) leading to the estimate

$$\hat{\pi}(\boldsymbol{\sigma}^* | \boldsymbol{y}, \mathcal{M}_1) = \frac{M^{-1} \sum_{g=1}^{M} \alpha(\boldsymbol{\sigma}^{(g)}, \boldsymbol{\sigma}^* | \boldsymbol{y}, \boldsymbol{\beta}^{(g)}, \{\boldsymbol{z}_i^{(g)}\})\, q(\boldsymbol{\sigma}^* | \boldsymbol{y}, \boldsymbol{\beta}^{(g)}, \{\boldsymbol{z}_i^{(g)}\})}{J^{-1} \sum_{j=1}^{J} \alpha(\boldsymbol{\sigma}^{(j)} | \boldsymbol{y}, \boldsymbol{\beta}^{(j)}, \{\boldsymbol{z}_i^{(j)}\})}, \tag{43}$$

where α is the probability of move defined in Algorithm 17, $\{\boldsymbol{\beta}^{(g)}, \{\boldsymbol{z}_i^{(g)}\}, \boldsymbol{\sigma}^{(g)}\}$ are values drawn from the full MCMC run and the values $\{\boldsymbol{\beta}^{(j)}, \{\boldsymbol{z}_i^{(j)}\}, \boldsymbol{\sigma}^{(j)}\}$ in the denominator are from a reduced run consisting of the densities

$$\pi(\boldsymbol{\beta} | \boldsymbol{y}, \mathcal{M}_1, \{z_i\}, \boldsymbol{\Sigma}^*); \quad \pi(\{z_i\} | \boldsymbol{y}, \mathcal{M}_1, \boldsymbol{\beta}, \boldsymbol{\Sigma}^*), \tag{44}$$

after $\boldsymbol{\Sigma}$ is fixed at $\boldsymbol{\Sigma}^*$. In particular, the draws for the denominator are from the distributions

$$\begin{aligned} \boldsymbol{\beta}^{(j)}, \boldsymbol{z}^{(j)} &\sim \pi(\boldsymbol{\beta}, \boldsymbol{z} | \boldsymbol{y}, \mathcal{M}_1, \boldsymbol{\Sigma}^*), \\ \boldsymbol{\sigma}^{(j)} &\sim q(\boldsymbol{\sigma}^*, \boldsymbol{\sigma} | \boldsymbol{y}, \mathcal{M}_1, \boldsymbol{\beta}^{(j)}, \boldsymbol{z}^{(j)}), \quad j \leqslant J. \end{aligned}$$

The sampled variates $\{\boldsymbol{\beta}^{(j)}, \boldsymbol{z}^{(j)}\}$ from this reduced run are also used to estimate the second ordinate as

$$\hat{\pi}(\boldsymbol{\beta}^* | \boldsymbol{y}, \mathcal{M}_1, \boldsymbol{\Sigma}^*) = M^{-1} \sum_{j=1}^{M} \phi_J(\boldsymbol{\beta}^* | \hat{\boldsymbol{\beta}}^{(j)}, \boldsymbol{B}_n^*), \tag{45}$$

where $\hat{\boldsymbol{\beta}}^{(j)} = \boldsymbol{B}_n(\boldsymbol{B}_0^{-1}\boldsymbol{\beta}_0 + \sum_{i=1}^{n} \boldsymbol{X}_i' \boldsymbol{\Sigma}^{*-1} \boldsymbol{z}_i^{(j)})$ and $\boldsymbol{B}_n^* = (\boldsymbol{B}_0^{-1} + \sum_{i=1}^{n} \boldsymbol{X}_i' \boldsymbol{\Sigma}^{*-1} \boldsymbol{X}_i)^{-1}$. It should be noted that estimates of *both* ordinates are available at the conclusion of the single reduced run.

Table 3
Log-likelihood and log marginal likelihood by the Chib method of three models fit to the Ohio wheeze data[1]

	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$
$\ln f(y \mid \mathcal{M}, \theta^*)$	−795.1869	−798.5567	−804.4102
$\ln m(y \mid \mathcal{M})$	−823.9188	−818.009	−824.0001

[1] $\mathcal{M}_1$, MVP with unrestricted correlations; $\mathcal{M}_2$, MVP with an equicorrelated correlation; $\mathcal{M}_3$, MVP with Toeplitz correlation structure.

The marginal likelihood computation is completed by evaluating the likelihood function at the point $(\boldsymbol{\beta}^*, \boldsymbol{\Sigma}^*)$ by the Geweke–Hajivassiliou–Keane method. The resulting marginal likelihoods of the three alternative models are reported in Table 3. On the basis of these marginal likelihoods we conclude that the data tend to support the MVP model with equicorrelated correlations.

10.3. Model space-parameter space MCMC algorithms

When one is presented with a large collection of candidate models $\{\mathcal{M}_1, \ldots, \mathcal{M}_K\}$, each with parameters $\boldsymbol{\theta}_k \in B_k \subseteq \Re^{d_k}$, direct fitting of each model to find the marginal likelihood can be computationally expensive. In such cases it may be more fruitful to utilize model space-parameter space MCMC algorithms that eschew direct fitting of each model for an alternative simulation of a "mega model" where a model index random variable, denoted as $\mathcal{M}$, taking values on the integers from 1 to K, is sampled in tandem with the parameters. The posterior distribution of $\mathcal{M}$ is then computed as the frequency of times each model is visited.

In this section we discuss two general model space-parameter space algorithms that have been proposed in the literature. These are the algorithms of Carlin and Chib (1995) and the reversible jump method of Green (1995).

To explain the Carlin and Chib (1995) algorithm, write $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K\}$ and assume that each model is defined by the likelihood $f(\boldsymbol{y}|\boldsymbol{\theta}_k, \mathcal{M} = k)$ and (proper) priors $p(\boldsymbol{\theta}_k|\mathcal{M} = k)$. Note that each model is non-nested. Now by the law of total probability the joint distribution of the data, the parameters and the model index is given by

$$f(\boldsymbol{y}, \boldsymbol{\theta}, \mathcal{M} = k) = f(\boldsymbol{y}|\boldsymbol{\theta}_k, \mathcal{M} = k)\, p(\boldsymbol{\theta}_k|\mathcal{M} = k)\, p(\boldsymbol{\theta}_{-k}|\boldsymbol{\theta}_k, \mathcal{M} = k) \Pr(\mathcal{M} = k). \tag{46}$$

Thus, in addition to the usual inputs, the joint probability model requires the specification of the densities $\{p(\boldsymbol{\theta}_{-k}|\boldsymbol{\theta}_k, \mathcal{M} = k),\ k \leqslant K\}$. These are called *pseudo priors* or *linking densities* and are necessary to complete the probability model but play no role in determining the marginal likelihood of $\mathcal{M} = k$ since

$$m(\boldsymbol{y}, \mathcal{M} = k) = \int f(\boldsymbol{y}, \boldsymbol{\theta}, \mathcal{M} = k)\, \mathrm{d}\boldsymbol{\theta},$$

regardless of what pseudo priors are chosen. Hence, the linking densities may be chosen in any convenient way that promotes the working of the MCMC sampling procedure. The goal now is to sample the posterior distribution on model space and parameter space

$$\pi(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K, \mathcal{M}|\boldsymbol{y}) \propto f(\boldsymbol{y}, \boldsymbol{\theta}, \mathcal{M}),$$

by MCMC methods.

Algorithm 18: Model space MCMC

```
(1) Sample
        $\boldsymbol{\theta}_k \sim \pi(\boldsymbol{\theta}_k|\boldsymbol{y}, \mathcal{M} = k) \propto f(\boldsymbol{y}|\boldsymbol{\theta}_k, \mathcal{M} = k)\,\pi(\boldsymbol{\theta}_k|\mathcal{M} = k), \quad \mathcal{M} = k,$
        $\boldsymbol{\theta}_{-k} \sim p(\boldsymbol{\theta}_{-k}|\boldsymbol{\theta}_k, \mathcal{M} = k), \quad \mathcal{M} \neq k.$
(2) Model jump
    (a) Calculate
        $p_k = \frac{f(\boldsymbol{y}|\boldsymbol{\theta}_k, \mathcal{M} = k)p(\boldsymbol{\theta}_k|\mathcal{M} = k)p(\boldsymbol{\theta}_{-k}|\boldsymbol{\theta}_k, \mathcal{M} = k)\Pr(\mathcal{M} = k)}{\sum_{l=1}^{K} f(\boldsymbol{y}|\boldsymbol{\theta}_l, \mathcal{M} = l)p(\boldsymbol{\theta}_l|\mathcal{M} = l)p(\boldsymbol{\theta}_{-l}|\boldsymbol{\theta}_l, \mathcal{M} = l)\Pr(M = l)}, k \leqslant K.$
    (b) Sample
        $\mathcal{M} \sim \{p_1, \ldots, p_K\}.$
(3) Goto 1.
```

Thus, when $\mathcal{M} = k$, we sample $\boldsymbol{\theta}_k$ from its full conditional distribution and the remaining parameters from their pseudo priors and the model index is sampled from the a discrete point distribution with probabilities $\{p_k\}$.

Algorithm 18 is conceptually quite simple and can be used without any difficulties when the number of models under consideration is small. When K is large, however, the specification of the pseudo priors and the requisite generation of each $\boldsymbol{\theta}_k$ within each cycle of the MCMC algorithm can be a computational burden. We also mention that the pseudo priors should be chosen to be close to the model specific posterior distributions. To understand the rationale for this recommendation suppose that the pseudo priors can be set exactly equal to the model specific posterior distributions as

$$p(\boldsymbol{\theta}_{-k}|\boldsymbol{\theta}_k, \mathcal{M} = k) = \prod_{l \neq k} \pi(\boldsymbol{\theta}_l|\boldsymbol{y}, \mathcal{M} = l).$$

Substituting this choice into the equation of p_k and simplifying we get

$$p_k = \frac{m(\boldsymbol{y}|\mathcal{M} = k)\Pr(\mathcal{M} = k)}{\sum_{l=1}^{K} m(\boldsymbol{y}|\mathcal{M} = l)\Pr(\mathcal{M} = l)}, \tag{47}$$

which is $\Pr(\mathcal{M} = k|\boldsymbol{y})$. Therefore, under this choice of pseudo priors, the Carlin–Chib algorithm generates the model move at each iteration of the sampling according to their posterior probabilities, without any required burn-in. Thus, by utilizing pseudo priors that are close to the model specific posterior distributions one promotes mixing on model space and more rapid convergence to the invariant target distribution $\pi(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_K, \mathcal{M}|\boldsymbol{y})$.

Another point in connection with the above algorithm is that the joint distribution over parameter space and model space can be sampled by the M–H algorithm. For example, Dellaportas, Forster and Ntzoufras (1998) suggest that the discrete conditional distribution on the models be sampled by M–H algorithm in order to avoid the calculation of the denominator of p_k. Godsill (1998) considers the sampling of the entire joint distribution in Equation (46) by the M–H algorithm. Suppose that the proposal density on the joint space is specified as

$$q\left\{(\mathcal{M} = k, \boldsymbol{\theta}_k, \boldsymbol{\theta}_{-k}), (\mathcal{M} = k', \boldsymbol{\theta}'_{k'}, \boldsymbol{\theta}'_{-k'})\right\} = q_1(k, k')\, q_2(\boldsymbol{\theta}_k, \boldsymbol{\theta}'_{k'}|k, k')\, p(\boldsymbol{\theta}'_{-k'}|\boldsymbol{\theta}'_{k'}, \mathcal{M} = k'), \tag{48}$$

where the pseudo prior is the proposal density of the parameters $\boldsymbol{\theta}_{-k'}$ not in the proposed model k'. It is important that q_1 not depend on the current value $(\boldsymbol{\theta}_k, \boldsymbol{\theta}_{-k})$ and that q_2 not depend on the current value of $\boldsymbol{\theta}_{k'}$ in the model being proposed. Then, the probability of move from $(\mathcal{M} = k, \boldsymbol{\theta}_k, \boldsymbol{\theta}_{-k})$ to $(\mathcal{M} = k', \boldsymbol{\theta}'_{k'}, \boldsymbol{\theta}'_{-k'})$ in the M–H step, after substitutions and cancellations, reduces to

$$\min\left\{1, \frac{f(\boldsymbol{y}|\boldsymbol{\theta}'_{k'}, \mathcal{M} = k')\, p(\boldsymbol{\theta}'_{k'}|\mathcal{M} = k')\, \Pr(\mathcal{M} = k')}{f(\boldsymbol{y}|\boldsymbol{\theta}_k, \mathcal{M} = k)\, p(\boldsymbol{\theta}_k|\mathcal{M} = k)\, \Pr(\mathcal{M} = k)}\, \frac{q_1(k', k)\, q_2(\boldsymbol{\theta}'_{k'}, \boldsymbol{\theta}_k|k, k')}{q_1(k, k')\, q_2(\boldsymbol{\theta}_k, \boldsymbol{\theta}'_{k'}|k, k')}\right\}, \tag{49}$$

which is completely independent of the pseudo priors. Thus, the sampling, or specification, of pseudo priors is not required in this version of the algorithm but the requirement that the parameters of each model be proposed in one block rules out many important problems.

We now turn to the reversible jump algorithm of Green (1995) which is designed primarily for *nested* models. In this algorithm, model space and parameter space moves from the current point $(\mathcal{M} = k, \boldsymbol{\theta}_k)$ to a new point $(\mathcal{M} = k', \boldsymbol{\theta}'_{k'})$ are made by a Metropolis–Hastings step in conjunction with a dimension matching condition to ensure that the resulting Markov chain is reversible. An application of the reversible jump method to choosing the number of components in a finite mixture of distribution model is provided by Richardson and Green (1997). The parameter space in this method is based on the *union* of the parameter spaces B_k. To describe the algorithm we let q denote a discrete mass function that gives the probability of each possible model given the current model and we let u' denote an increment/decrement random variable that takes one from the current point $\boldsymbol{\theta}_k$ to the new point $\boldsymbol{\theta}'_{k'}$.

Algorithm 19: Reversible jump model space MCMC

```
(1) Propose a new model $k'$

        $k' \sim q_1(k, k')$.
```

```
(2) Dimension matching
    (a) Propose
            $u' \sim q_2(u'|\boldsymbol{\theta}_k, k, k')$.
    (b) Set
            $(\boldsymbol{\theta}'_{k'}, u) = g_{k,k'}(\boldsymbol{\theta}_k, u')$,
        where $g_{k,k'}$ is a bijection between $(\boldsymbol{\theta}'_{k'}, u)$ and $(\boldsymbol{\theta}_k, u')$ and $\dim(\boldsymbol{\theta}_k)$
        $+ \dim(u') = \dim(\boldsymbol{\theta}'_{k'}) + \dim(u)$.
(3) M-H
    (a) Calculate
            $\alpha = \min\left\{1, \frac{f(\boldsymbol{y}|\boldsymbol{\theta}'_{k'}, \mathcal{M}=k')p(\boldsymbol{\theta}'_{k'}|\mathcal{M}=k')\Pr(\mathcal{M}=k')}{f(\boldsymbol{y}|\boldsymbol{\theta}_k, \mathcal{M}=k)p(\boldsymbol{\theta}_k|\mathcal{M}=k)\Pr(\mathcal{M}=k)} \frac{q_1(k',k)q_2(u|\boldsymbol{\theta}_k,k,k')}{q_2(k,k')q_2(u'|\boldsymbol{\theta}_k,k,k')} \cdot J\right\}$,
        where
            $J = \left|\frac{\partial g_{k,k'}(\boldsymbol{\theta}_k, u')}{\partial(\boldsymbol{\theta}_k, u')}\right|$.
    (b) Move to $(k'; \boldsymbol{\theta}'_{k'}, u')$ with probability $\alpha$.
(4) Goto 1.
```

In the reversible jump method most of the tuning is in the specification of the proposal distribution q_2; a different proposal distribution is required if k' is a model with more parameters than model k than for the case when model k' has fewer parameters. This is the reason for the dependence of q_2 on not just $\boldsymbol{\theta}_k$ but also on (k, k'). In addition, the algorithm as stated by Green (1995) is designed for the situation where the competing models are nested and obtained by the removal or addition of different parameters, as for example in a variable selection problem.

10.4. Variable selection

Model space MCMC methods described above can be specialized to the problem of variable selection in regression. We first focus on this problem in the context of linear regression models with conjugate priors before discussing a more general situation.

Consider then the question of building a multiple regression model for a vector of n observations $\boldsymbol{y}$ in terms of a given set of covariates $\boldsymbol{X} = \{x_1, \ldots, x_p\}$. The goal is to find the "best" model of the form

$$\mathcal{M}_k\colon \boldsymbol{y} = \boldsymbol{X}_k \boldsymbol{\beta}_k + \sigma\varepsilon,$$

where $\boldsymbol{X}_k$ is a $n \times d_k$ matrix composed of some or all variables from $\boldsymbol{X}$, σ^2 is a variance parameter and ε is $\mathcal{N}(0, \boldsymbol{I}_n)$. Under the assumption that any subset of the variables in

$\boldsymbol{X}$ can be used to form $\boldsymbol{X}_k$ it follows that the number of possible models is given by $K = 2^p$, which is a large number even if p is as small as fifteen. Thus, unless p is small, when the marginal likelihoods can be computed for each possible $\boldsymbol{X}_k$, it is helpful to use simulation-based methods that traverse the space of possible models to determine the subsets that are most supported by the data.

Raftery, Madigan and Hoeting (1997) develop one approach that is based on the use of conjugate priors. Let the parameters $\boldsymbol{\theta}_k = (\boldsymbol{\beta}_k, \sigma^2)$ of model $\mathcal{M}_k$ follow the conjugate prior distributions

$$\boldsymbol{\beta}_k|\mathcal{M} = k, \sigma^2 \sim \mathcal{N}_{d_k}(\boldsymbol{0}, \sigma^2\boldsymbol{B}_{0k});\ \sigma^2|\mathcal{M} = k \sim \mathcal{IG}\left(\frac{\nu_0}{2}, \frac{\delta_0}{2}\right), \tag{50}$$

which implies after some algebra that the marginal likelihood of $\mathcal{M}_k$ is

$$m(\boldsymbol{y}|\mathcal{M} = k) = \frac{\Gamma\{(\nu_0 + n)/2\}}{\Gamma(\nu_0/2)(\delta_0\pi)^{n/2}}\,|\boldsymbol{B}_k|^{1/2} \times \left(1 + \frac{1}{\delta_0}\boldsymbol{y}'\boldsymbol{B}_k\boldsymbol{y}\right)^{-(n+\nu_0)/2},$$

where

$$\boldsymbol{B}_k = \boldsymbol{I}_n - \boldsymbol{X}_k(\boldsymbol{B}_{0k}^{-1} + \boldsymbol{X}_k'\boldsymbol{X}_k)^{-1}\boldsymbol{X}_k'.$$

Raftery, Madigan and Hoeting (1997) specify a MCMC chain to sample model space in which the target distribution is the *univariate* discrete distribution with probabilities

$$\Pr(\mathcal{M} = k|\,\boldsymbol{y}) = p_k \propto m(\boldsymbol{y}|\mathcal{M} = k)\Pr(\mathcal{M} = k), \quad k \leqslant K. \tag{51}$$

Although this distribution can in principle be normalized, the normalization constant is computationally expensive to calculate when K is large (but one can argue that expending the necessary computational effort is always desirable). This motivates the sampling of Equation (51) by the Metropolis–Hastings algorithm. For each model $\mathcal{M} = k$ define a neighborhood nbd($\mathcal{M} = k$) which consists of the model $\mathcal{M} = k$ and models with either one more variable or one fewer variable than $\mathcal{M} = k$. Define a transition matrix $q_1(k, k')$ which puts uniform probability over models k' that are in nbd($\mathcal{M} = k$) and zero probability for all other models. Given that the chain is currently at the point ($\mathcal{M} = k$) a move to the proposed model k' is made with probability

$$\min\left\{\frac{m(\boldsymbol{y}|\mathcal{M} = k')\Pr(\mathcal{M} = k')}{m(\boldsymbol{y}|\mathcal{M} = k)\Pr(\mathcal{M} = k)}\frac{q_1(k', k)}{q_1(k, k')}, 1\right\}. \tag{52}$$

If the proposed move is rejected the chain stays at $\mathcal{M} = k$.

When conjugate priors are not assumed for $\boldsymbol{\theta}_k$, or when the model is more complicated than multiple regression, it is not possible to find the marginal likelihood of each model analytically. It then becomes necessary to sample both the parameters and the model index jointly as in the general model space-parameter space algorithms mentioned above. The approaches that have been developed for this case, however, treat the various models as nested.

Suppose that the coefficients attached to the p possible covariates in the model are denoted by $\eta = \{\eta_1, \ldots, \eta_p\}$, where any common noise variances or other common parameters are suppressed from the notation and the discussion. Now associate with each coefficient η_j an indicator variable δ_j which takes the value one if the coefficient is in the model and the value zero otherwise and let η_δ denote the set of active η_j's given a configuration $\boldsymbol{\delta}$ and let $\eta_{-\boldsymbol{\delta}}$ denote the complementary η_j's. For example, if $p = 5$ and $\boldsymbol{\delta} = \{1, 0, 0, 1, 1\}$, then $\eta_{\boldsymbol{\delta}} = \{\eta_1, \eta_4, \eta_5\}$ and $\eta_{-\boldsymbol{\delta}} = \{\eta_2, \eta_3\}$. A variable selection MCMC algorithm can now be developed by sampling the joint posterior distribution $\pi(\delta_1, \eta_1, \ldots, \delta_p, \eta_p | \boldsymbol{y})$. Particular implementations representing different blocking schemes to sample this joint distribution are discussed by Kuo and Mallick (1998), Geweke (1996) and Smith and Kohn (1996). For example, in the algorithm of Kuo and Mallick (1998), the posterior distribution is sampled by recursively simulating the $\{\eta_1, \ldots, \eta_p\}$ from the distributions

$$\eta_j \sim \pi(\eta_j | \boldsymbol{y}, \eta_{-j}, \boldsymbol{\delta}) \propto \begin{cases} f(\boldsymbol{y} | \eta_{\boldsymbol{\delta}}, \boldsymbol{\delta}) p(\eta_{\boldsymbol{\delta}} | \boldsymbol{\delta}) & \text{if } \delta_j = 1, \\ p(\eta_j | \eta_{-j}, \boldsymbol{\delta}) & \text{if } \delta_j = 0, \end{cases}$$

where $p(\eta_j | \eta_{-j}, \boldsymbol{\delta})$ is a pseudo prior because it represents the distribution of η_j when η_j is not in the current configuration. Next, the variable indicators $\{\delta_1, \ldots, \delta_p\}$ are sampled one at a time from the two point mass function

$$\delta_j \sim \Pr(\delta_j | \boldsymbol{y}, \eta_{-j}, \boldsymbol{\delta}_{-j}) \propto f(\boldsymbol{y} | \eta_{\boldsymbol{\delta}}, \boldsymbol{\delta})\, p(\eta_{\boldsymbol{\delta}} | \boldsymbol{\delta})\, p(\eta_{-\boldsymbol{\delta}} | \eta_{\boldsymbol{\delta}}, \boldsymbol{\delta})\, p(\delta_j),$$

where $p(\eta_{-\boldsymbol{\delta}} | \eta_{\boldsymbol{\delta}}, \boldsymbol{\delta})$ is the pseudo prior. These two steps are iterated. Procedures to sample (δ_j, η_j) in one block given all the other blocks are presented by Geweke (1996) and Smith and Kohn (1996).

George and McCulloch (1993, 1997) develop an important alternative simulation-based approach for the variable selection problem that has been extensively studied and refined. In their approach, the variable selection problem is cast in terms of a hierarchical model of the type

$$\begin{aligned} &\boldsymbol{y} \sim \boldsymbol{X\beta} + \sigma\varepsilon, \quad \boldsymbol{\beta}_j | \gamma_j \sim (1 - \gamma_j) N(0, \tau_j^2) + \gamma_j N(0, c_j^2 \tau_j^2), \\ &\quad \Pr(\gamma_j = 1) = 1 - \Pr(\gamma_j = 0) = p_j, \end{aligned}$$

where τ_j^2 is a small positive number and c_j a large positive number. In this specification each component of $\boldsymbol{\beta}$ is assumed to come from a mixture of two normal distributions such that $\gamma_j = 0$ corresponds to the case where $\boldsymbol{\beta}_j$ can be assumed to be zero. It should be noted that in this framework a particular covariate is never strictly removed from the model; exclusion from the model corresponds to a high posterior probability of the event that $\gamma_j = 0$. George and McCulloch (1993) sample the posterior distribution of $(\boldsymbol{\beta}, \{\gamma_j\})$ by the Gibbs sampling algorithm.

10.5. Remark

We conclude this discussion by pointing out that convergence checks of the Markov chain in model space algorithms is quite difficult and has not been satisfactorily addressed in the literature. When the model space is large, as for example in the variable selection problem, one cannot be sure that all models supported by the data have been visited according to their posterior probabilities. Of course if the model space is diminished to ensure better coverage of the various models it may happen that direct computation of the marginal likelihood becomes feasible, thereby removing any justification for considering a model space algorithm in the first place. This tension in the choice between direct computation and model space algorithms is real and cannot be adjudicated in the absence of a concrete problem.

11. MCMC methods in optimization problems

Suppose that we are given a particular function $h(\boldsymbol{\theta})$, say the log likelihood of a given model, and interest lies in the value of $\boldsymbol{\theta}$ that maximizes this function. In some cases, this optimization problem can be quite effectively solved by MCMC methods. One somewhat coarse possibility is to obtain draws $\{\boldsymbol{\theta}^{(j)}\}$ from a density proportional to $h(\boldsymbol{\theta})$ and to find the value of $\boldsymbol{\theta}$ that corresponds to the maximum of $\{h(\boldsymbol{\theta}^{(j)})\}$. Another more precise technique goes by the name of simulated annealing which appears in Metropolis et al. (1953) and is closely related to the Metropolis simulation method. In the simulated annealing method, which is most typically used to maximize a function on a finite but large set, one uses the Metropolis method to sample the distribution

$$\pi(\boldsymbol{\theta}) \propto \exp\{h(\boldsymbol{\theta})/T\},$$

where T is referred to as the temperature. The temperature variable is gradually reduced as the sampling proceeds [for example, see Geman and Geman (1984)]. It can be shown that in the finite case, the values of $\boldsymbol{\theta}$ produced by the simulated annealing method concentrate around the local maximum of the function $h(\boldsymbol{\theta})$.

Another method of interest is a MCMC version of the EM algorithm which can be used to find the maximum likelihood estimate in certain situations. Suppose that $\boldsymbol{z}$ represents missing data and $f(\boldsymbol{y}|\mathcal{M}, \boldsymbol{\theta})$ denotes the likelihood function. Also suppose that

$$f(\boldsymbol{y}|\mathcal{M}, \boldsymbol{\theta}) = \int f(\boldsymbol{y}, \boldsymbol{z}|\mathcal{M}, \boldsymbol{\theta})\,\mathrm{d}\boldsymbol{z},$$

is difficult to compute but that the complete data likelihood $f(\boldsymbol{y}, \boldsymbol{z}|\mathcal{M}, \boldsymbol{\theta})$ is available, as in the models with a missing data structure in Section 8. For this problem, the standard EM algorithm [Dempster, Laird and Rubin (1977)] requires the recursive implementation of two steps: the expectation or E-step and the maximization or M-step. In the E-step, given the current guess of the maximizer $\boldsymbol{\theta}^{(j)}$, one computes

$$Q(\boldsymbol{\theta}^{(j)}, \boldsymbol{\theta}) = \int \ln f(\boldsymbol{y}, \boldsymbol{z} | \mathcal{M}, \boldsymbol{\theta}) f(\boldsymbol{z} | \boldsymbol{y}, \mathcal{M}, \boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{z},$$

while in the M-step the Q function is maximized to obtain a revised guess of the maximizer, i.e.,

$$\boldsymbol{\theta}^{(j+1)} = \arg\max_{\boldsymbol{\theta}} \, Q(\boldsymbol{\theta}^{(j)}, \boldsymbol{\theta}).$$

Wu (1983) has shown that under regularity conditions the sequence of values $\{\boldsymbol{\theta}^{(j)}\}$ generated by these steps converges to the maximizer of the function $f(\boldsymbol{y} | \mathcal{M}, \boldsymbol{\theta})$.

The MCEM algorithm is a variant of the EM algorithm, proposed by Wei and Tanner (1990b), in which the E-step, which is often intractable, is computed by Monte Carlo averaging over values of $\boldsymbol{z}$ drawn from $f(\boldsymbol{z} | \boldsymbol{y}, \mathcal{M}, \boldsymbol{\theta})$, which in the MCMC context is the full conditional distribution of the latent data. Then, the revised value of $\boldsymbol{\theta}$ is obtained by maximizing the Monte Carlo estimate of the Q function. Specifically, the MCEM algorithm is defined by iterating on the following steps:

$$\hat{Q}_M(\boldsymbol{\theta}^{(j)}, \boldsymbol{\theta}) = M^{-1} \sum_{j=1}^{M} \ln f(\boldsymbol{y}, \boldsymbol{z}^{(j)} | \mathcal{M}, \boldsymbol{\theta}),$$

$$\boldsymbol{z}^{(j)} \sim f(\boldsymbol{z} | \boldsymbol{y}, \mathcal{M}, \boldsymbol{\theta}), \quad \boldsymbol{\theta}^{(j+1)} = \arg\max_{\boldsymbol{\theta}} \, \hat{Q}(\boldsymbol{\theta}^{(j)}, \boldsymbol{\theta}).$$

As suggested by Wei and Tanner (1990b), these iterations are started with a small value of M that is increased as the maximizer is approached. One point to note is that in general, the MCEM algorithm, similar to the EM algorithm, can be slow to converge to the mode but it should be possible to adapt the ideas described in Liu, Rubin and Wu (1998) to address this problem. Another point to note is that the computation of the $\hat{Q}_M$ function can be expensive when M is large. Despite these potential difficulties, a number of applications of the MCEM algorithm have now appeared in the literature. These include Chan and Ledolter (1995), Chib (1996, 1998), Chib and Greenberg (1998), Chib, Greenberg and Winkelmann (1998) and Booth and Hobert (1999).

Given the modal value $\hat{\theta}$, the standard errors of the MLE are obtained by the formula of Louis (1982). In particular, the observed information matrix is given by

$$-E\left\{\frac{\partial^2 \ln f(\boldsymbol{y}, \boldsymbol{z} | \mathcal{M}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right\} - \mathrm{Var}\left\{\frac{\partial \ln f(\boldsymbol{y}, \boldsymbol{z} | \mathcal{M}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right\},$$

where the expectation and variance are with respect to the distribution $\boldsymbol{z}|\boldsymbol{y}, \mathcal{M}, \hat{\theta}$. This expression is estimated by taking an additional J draws $\{\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(J)}\}$ from $\boldsymbol{z}|\boldsymbol{y}, \mathcal{M}, \hat{\theta}$ and computing

$$
\begin{aligned}
&-J^{-1}\sum_{k=1}^{J}\frac{\partial^2 \ln f(\boldsymbol{y}, \boldsymbol{z}^{(k)}|\mathcal{M}, \boldsymbol{\theta}^*)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} \\
&-J^{-1}\sum_{k=1}^{J}\left(\frac{\partial \ln f(\boldsymbol{y}, \boldsymbol{z}^{(k)}|\mathcal{M}, \boldsymbol{\theta}^*)}{\partial\boldsymbol{\theta}} - \boldsymbol{m}\right)\left(\frac{\partial \ln f(\boldsymbol{y}, \boldsymbol{z}^{(k)}|\mathcal{M}, \boldsymbol{\theta}^*)}{\partial\boldsymbol{\theta}} - \boldsymbol{m}\right)',
\end{aligned}
$$

where

$$
\boldsymbol{m} = J^{-1}\sum_{k=1}^{J}\frac{\partial \ln f(\boldsymbol{y}, \boldsymbol{z}^{(k)}|\mathcal{M}, \hat{\theta})}{\partial\boldsymbol{\theta}}.
$$

Standard errors are equal to the square roots of the diagonal elements of the inverse of the estimated information matrix.

12. Concluding remarks

In this survey we have provided an outline of Markov chain Monte Carlo methods with emphasis on techniques that prove useful in Bayesian statistical inference. Further developments of these methods continue to occur but the ideas and details presented in this survey should provide a reasonable starting point to understand the current and emerging literature. Two recent developments are the slice sampling method discussed by Mira and Tierney (1998), Damien et al. (1999) and Roberts and Rosenthal (1999) and the perfect sampling method proposed by Propp and Wilson (1996). The slice sampling method is based on the introduction of auxiliary uniform random variables to simplify the sampling and improve mixing while the perfect sampling method uses Markov chain coupling to generate an exact draw from the target distribution. These methods are in their infancy and can be currently applied only under rather restrictive assumptions on the target distribution but it is possible that more general versions of these methods will eventually become available.

Other interesting developments are now occurring in the field of applied Bayesian inference as practical problems are being addressed by the methods summarized in this survey. These applications are appearing at a steady rate in various areas. For example, a partial list of fields and papers within fields include: biostatistical time series analysis [West, Prado and Krystal (1999)]; economics [Chamberlain and Hirano (1997), Filardo and Gordon (1998), Gawande (1998), Lancaster (1997), Li (1998), Kiefer and Steel (1998), Kim and Nelson (1999), Koop and Potter (1999), Martin (1999), Paap and van Dijk (1999), So, Lam and Li (1998)]; finance [Jones (1999), Pastor and Stambaugh

(1999)]; marketing [Allenby, Leone and Jen (1999), Bradlow and Zaslavsky (1999), Manchanda, Ansari and Gupta (1999), Montgomery and Rossi (1999), Young, DeSarbo and Morwitz (1998)]; political science [King, Rosen and Tanner (1999), Quinn, Martin and Whitford (1999), Smith (1999)]; and many others.

One can claim that with the ever increasing power of computing hardware, and the experience of the past ten years, the future of simulation-based inference using MCMC methods is secure.

References

Albert, J. (1993), "Teaching Bayesian statistics using sampling methods and MINITAB", American Statistician 47:182–191.

Albert, J., and S. Chib (1993a), "Bayesian analysis of binary and polychotomous response data", Journal of the American Statistical Association 88:669–679.

Albert, J., and S. Chib (1993b), "Bayes inference via Gibbs sampling of autoregressive time series subject to Markov mean and variance shifts", Journal of Business and Economic Statistics 11:1–15.

Albert, J., and S. Chib (1995), "Bayesian residual analysis for binary response models", Biometrika 82:747–759.

Albert, J., and S. Chib (1996), "Computation in Bayesian Econometrics: An Introduction to Markov Chain Monte Carlo", in: T. Fomby and R.C. Hill, eds., Advances in Econometrics, Vol. 11A (Jai Press, Greenwich, CT) 3–24.

Albert, J., and S. Chib (1997), "Bayesian tests and model diagnostics in conditionally independent hierarchical models", Journal of the American Statistical Association 92:916–925.

Albert, J., and S. Chib (1998), "Sequential Ordinal Modeling with Applications to Survival Data". Biometrics, in press.

Allenby, G.M., R.P. Leone and L. Jen (1999), "A dynamic model of purchase timing with application to direct marketing", Journal of the American Statistical Association 94:365–374.

Bennett, C.H. (1976), "Efficient estimation of free energy differences from Monte Carlo data", Journal of Computational Physics 22:245–268.

Berger, J.O. (1985), Statistical Decision Theory and Bayesian Analysis, 2nd edition (Springer, New York).

Bernardo, J.M., and A.F.M. Smith (1994), Bayesian Theory (Wiley, New York).

Besag, J. (1974), "Spatial interaction and the statistical analysis of lattice systems (with discussion)", Journal of the Royal Statistical Society B 36:192–236.

Besag, J., E. Green, D. Higdon and K.L. Mengersen (1995), "Bayesian computation and stochastic systems (with discussion)", Statistical Science 10:3–66.

Best, N.G., M.K. Cowles and S.K. Vines (1995), "CODA: convergence diagnostics and output analysis software for Gibbs sampling", Technical report (Cambridge MRC Biostatistics Unit).

Billio, M., A. Monfort and C.P. Robert (1999), "Bayesian estimation of switching ARMA models", Journal of Econometrics 93:229–255.

Blattberg, R.C., and E.I. George (1991), "Shrinkage estimation of price and promotional elasticities: seemingly unrelated equations", Journal of the American Statistical Association 86:304–315.

Booth, J.G., and J.P. Hobert (1999), "Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm", Journal of the Royal Statistical Society B 61:265–285.

Bradlow, E., and A.M. Zaslavsky (1999), "A hierarchical latent variable model for ordinal data from a customer satisfaction survey with "no answer" responses", Journal of the American Statistical Association 94:43–52.

Brooks, S.P. (1998), "Markov chain Monte Carlo and its application", Statistician 47:69–100.

Brooks, S.P., P. Dellaportas and G.O. Roberts (1997), "A total variation method for diagnosing convergence of MCMC algorithms", Journal of Computational and Graphical Statistics 6:251–265.

Carlin, B., A.E. Gelfand and A.F.M. Smith (1992), "Hierarchical Bayesian analysis of changepoint problems", Applied Statistics 41:389–405.

Carlin, B.P., and S. Chib (1995), "Bayesian model choice via Markov Chain Monte Carlo methods", Journal of the Royal Statistical Society B 57:473–484.

Carlin, B.P., and T.A. Louis (2000), Bayes and Empirical Bayes Methods for Data Analysis, 2nd Edition (Chapman and Hall, London).

Carlin, B.P., and N.G. Polson (1991), "Inference for non-conjugate Bayesian models using the Gibbs sampler", Canadian Journal of Statistics 19:399–405.

Carlin, B.P., N.G. Polson and D.S. Stoffer (1992), "A Monte Carlo approach to nonnormal and nonlinear state-space modeling", Journal of the American Statistical Association 87:493–500.

Carter, C., and R. Kohn (1994), "On Gibbs sampling for state space models", Biometrika 81:541–553.

Carter, C., and R. Kohn (1996), "Markov chain Monte Carlo for conditionally Gaussian state space models", Biometrika 83:589–601.

Casella, G., and E.I. George (1992), "Explaining the Gibbs sampler", American Statistician 46:167–174.

Casella, G., and C.P. Robert (1996), "Rao-Blackwellization of sampling schemes", Biometrika 83:81–94.

Chamberlain, G., and K. Hirano (1997), "Predictive distributions based on longitudinal earnings data", Manuscript (Department of Economics, Harvard University).

Chan, K.S. (1993), "Asymptotic behavior of the Gibbs sampler", Journal of the American Statistical Association 88:320–326.

Chan, K.S., and C.J. Geyer (1994), "Discussion of Markov chains for exploring posterior distributions", Annals of Statistics 22:1747–1758.

Chan, K.S., and J. Ledolter (1995), "Monte Carlo EM estimation for time series models involving counts", Journal of the American Statistical Association 90:242–252.

Chen, M.-H. (1994), "Importance-weighted marginal Bayesian posterior density estimation", Journal of the American Statistical Association 89:818–824.

Chen, M.-H., and Q.-M. Shao (1997), "On Monte Carlo methods for estimating ratios of normalizing constants", Annals of Statistics 25:1563–1594.

Chen, M.-H., and Q.-M. Shao (1999), "Monte Carlo estimation of Bayesian credible and HPD intervals", Journal of Computational and Graphical Statistics 8:69–92.

Chib, S. (1992), "Bayes regression for the Tobit censored regression model", Journal of Econometrics 51:79–99.

Chib, S. (1993), "Bayes regression with autocorrelated errors: a Gibbs sampling approach", Journal of Econometrics 58:275–294.

Chib, S. (1995), "Marginal likelihood from the Gibbs output", Journal of the American Statistical Association 90:1313–1321.

Chib, S. (1996), "Calculating posterior distributions and modal estimates in Markov mixture models", Journal of Econometrics 75:79–97.

Chib, S. (1998), "Estimation and comparison of multiple change point models", Journal of Econometrics 86:221–241.

Chib, S., and B.P. Carlin (1999), "On MCMC sampling in hierarchical longitudinal models", Statistics and Computing 9:17–26.

Chib, S., and E. Greenberg (1994), "Bayes inference for regression models with ARMA(p,q) errors", Journal of Econometrics 64:183–206.

Chib, S., and E. Greenberg (1995a), "Understanding the Metropolis–Hastings algorithm", American Statistician 49:327–335.

Chib, S., and E. Greenberg (1995b), "Hierarchical analysis of SUR models with extensions to correlated serial errors and time-varying parameter models", Journal of Econometrics 68:339–360.

Chib, S., and E. Greenberg (1996), "Markov chain Monte Carlo simulation methods in econometrics", Econometric Theory 12:409–431.

Chib, S., and E. Greenberg (1998), "Analysis of multivariate probit models", Biometrika 85:347–361.

Chib, S., and B. Hamilton (2000), "Bayesian analysis of cross section and clustered data treatment models", Journal of Econometrics 97:25–50.

Chib, S., and I. Jeliazkov (2001), "Marginal likelihood from the Metropolis–Hastings output", Journal of the American Statistical Association 96:270–281.

Chib, S., E. Greenberg and R. Winkelmann (1998), "Posterior simulation and Bayes factors in panel count data models", Journal of Econometrics 86:33–54.

Chib, S., F. Nardari and N. Shephard (1998), "Markov Chain Monte Carlo analysis of generalized stochastic volatility models", Journal of Econometrics, under review.

Chib, S., F. Nardari and N. Shephard (1999), "Analysis of high dimensional multivariate stochastic volatility models", Technical report (John M. Olin School of Business, Washington University, St. Louis).

Chipman, H.A., E.I. George and R.E. McCulloch (1998), "Bayesian CART model search (with discussion)", Journal of the American Statistical Association 93:935–948.

Cowles, M.K. (1996), "Accelerating Monte Carlo Markov chain convergence for cumulative-link generalized linear models", Statistics and Computing 6:101–111.

Cowles, M.K., and B. Carlin (1996), "Markov chain Monte Carlo convergence diagnostics: a comparative review", Journal of the American Statistical Association 91:883–904.

Damien, P., J. Wakefield and S. Walker (1999), "Gibbs sampling for Bayesian nonconjugate and hierarchical models using auxiliary variables", Journal of the Royal Statistical Society B 61:331–344.

de Jong, P., and N. Shephard (1995), "The simulation smoother for time series models", Biometrika 82:339–350.

Dellaportas, P., and A.F.M. Smith (1993), "Bayesian inference for generalized linear and proportional hazards models via Gibbs sampling", Applied Statistics 42:443–459.

Dellaportas, P., J.J. Forster and I. Ntzoufras (1998), "On Bayesian model and variable selection using MCMC", Technical report (University of Economics and Business, Greece).

Dempster, A.P., N.M. Laird and D.B. Rubin (1977), "Maximum likelihood estimation from incomplete data via the EM algorithm", Journal of the Royal Statistical Society B 39:1–38.

Denison, D.G.T., B.K. Mallick and A.F.M. Smith (1998), "A Bayesian CART algorithm", Biometrika 85:363–377.

Devroye, L. (1985), Non-Uniform Random Variate Generation (Springer, New York).

DiCiccio, T.J., R.E. Kass, A.E. Raftery and L. Wasserman (1997), "Computing Bayes factors by combining simulation and asymptotic approximations", Journal of the American Statistical Association 92:903–915.

Diebolt, J., and C.P. Robert (1994), "Estimation of finite mixture distributions through Bayesian sampling", Journal of the Royal Statistical Society B 56:363–375.

Diggle, P., K.-Y. Liang and S.L. Zeger (1995), Analysis of Longitudinal Data (Oxford University Press, Oxford).

Elerian, O., S. Chib and N. Shephard (1999), "Likelihood inference for discretely observed nonlinear diffusions", Econometrica, in press.

Escobar, M.D., and M. West (1995), "Bayesian prediction and density estimation", Journal of the American Statistical Association 90:577–588.

Filardo, A.J., and S.F. Gordon (1998), "Business cycle durations", Journal of Econometrics 85:99–123.

Fruhwirth-Schnatter, S. (1994), "Data augmentation and dynamic linear models", Journal of Time Series Analysis 15:183–202.

Gammerman, D. (1997), Markov chain Monte Carlo: Stochastic Simulation for Bayesian Inference (Chapman and Hall, London).

Gammerman, D., and H.S. Migon (1993), "Dynamic hierarchical models", Journal of the Royal Statistical Society B 55:629–642.

Gawande, K. (1998), "Comparing theories of endogenous protection: Bayesian comparison of Tobit models using Gibbs sampling output", Review of Economics and Statistics 80:128–140.

Gelfand, A.E., and D. Dey (1994), "Bayesian model choice: asymptotics and exact calculations", Journal of the Royal Statistical Society B 56:501–514.

Gelfand, A.E., and A.F.M. Smith (1990), "Sampling-based approaches to calculating marginal densities", Journal of the American Statistical Association 85:398–409.

Gelfand, A.E., and A.F.M. Smith (1992), "Bayesian statistics without tears: a sampling–resampling perspective", American Statistician 46:84–88.

Gelfand, A.E., S. Hills, A. Racine-Poon and A.F.M. Smith (1990), "Illustration of Bayesian inference in normal data models using Gibbs sampling", Journal of the American Statistical Association 85:972–982.

Gelfand, A.E., S.K. Sahu and B.P. Carlin (1995), "Efficient parameterizations for normal linear mixed models", Biometrika 82:479–488.

Gelman, A., and D.B. Rubin (1992), "Inference from iterative simulation using multiple sequences", Statistical Science 4:457–472.

Gelman, A., X.L. Meng, H.S. Stern and D.B. Rubin (1995), Bayesian Data Analysis (Chapman and Hall, London).

Geman, S., and D. Geman (1984), "Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images", IEEE Transactions on Pattern Analysis and Machine Intelligence 12:609–628.

Gentle, J.E. (1998), Random Number Generation and Monte Carlo Methods (Springer, New York).

George, E.I., and R.E. McCulloch (1993), "Variable selection via Gibbs sampling", Journal of the American Statistical Association 88:881–889.

George, E.I., and R.E. McCulloch (1997), "Approaches to Bayesian variable selection", Statistica Sinica 7:339–373.

Geweke, J. (1989), "Bayesian inference in econometric models using Monte Carlo integration", Econometrica 57:1317–1340.

Geweke, J. (1991), "Efficient simulation from the multivariate normal and student-*t* distributions subject to linear constraints", in: E. Keramidas and S. Kaufman, eds., Computing Science and Statistics: Proceedings of the 23rd Symposium (Interface Foundation of North America) 571–578.

Geweke, J. (1992), "Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments", in: J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith, eds., Bayesian Statistics (Oxford University Press, New York) 169–193.

Geweke, J. (1996), "Variable selection and model comparison in regression", in: J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith, eds., Bayesian Statistics (Oxford University Press, New York) 609–620.

Geweke, J. (1997), "Posterior simulators in econometrics", in: D.M. Kreps and K.F. Wallis, eds., Advances in Economics and Econometrics: Theory and Applications, 7th World Congress (Cambridge University Press, Cambridge) 128–165.

Geyer, C. (1995), "Conditioning in Markov chain Monte Carlo", Journal of Computational and Graphical Statistics 4:148–154.

Geyer, C.J., and E.A. Thompson (1995), "Annealing Markov chain Monte Carlo with applications to ancestral inference", Journal of the American Statistical Association 90:909–920.

Ghysels, E., A.C. Harvey and E. Renault (1996), "Stochastic volatility", in: C.R. Rao and G.S. Maddala, eds., Statistical Methods in Finance (North-Holland, Amsterdam) 119–191.

Gilks, W.R., S. Richardson and D.J. Spiegelhalter (1996), Markov Chain Monte Carlo in Practice (Chapman and Hall, London).

Godsill, S.J. (1998), "On the relationship between model uncertainty methods", Technical report (Signal Processing Group, Cambridge University).

Green, P.E. (1995), "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination", Biometrika 82:711–732.

Hamilton, J.D. (1989), "A new approach to the economic analysis of nonstationary time series subject to changes in regime", Econometrica 57:357–384.

Hand, D.J., F. Daly, A.D. Lunn, K.J. McConway and E. Ostrowski (1994), A Handbook of Small Data Sets (Chapman and Hall, London).

Hastings, W.K. (1970), "Monte Carlo sampling methods using Markov chains and their applications", Biometrika 57:97–109.

Hills, S.E., and A.F.M. Smith (1992), "Parameterization issues in Bayesian inference", in: J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith, eds., Proceedings of the Fourth Valencia International Conference on Bayesian Statistics (Oxford University Press, New York) 641–649.

Jacquier, E., N.G. Polson and P.E. Rossi (1994), "Bayesian analysis of stochastic volatility models (with discussion)", Journal of Business and Economic Statistics 12:371–417.

Jeffreys, H. (1961), Theory of Probability, 3rd edition (Oxford University Press, New York).

Jones, C.S. (1999), "The dynamics of stochastic volatility", Manuscript (University of Rochester).

Kiefer, N.M., and M.F.J. Steel (1998), "Bayesian analysis of the prototypal search model", Journal of Business and Economic Statistics 16:178–186.

Kim, C.-J., and C.R. Nelson (1999), "Has the US become more stable? A Bayesian approach based on a Markov-switching model of business cycle", The Review of Economics and Statistics 81:608–616.

Kim, S., N. Shephard and S. Chib (1998), "Stochastic volatility: likelihood inference and comparison with ARCH models", Review of Economic Studies 65:361–393.

King, G., O. Rosen and M.A. Tanner (1999), "Binomial-beta hierarchical models for ecological inference", Sociological Method Research 28:61–90.

Kloek, T., and H.K. van Dijk (1978), "Bayesian estimates of equation system parameters: an application of integration by Monte Carlo", Econometrica 46:1–20.

Koop, G., and S.M. Potter (1999), "Bayes factors and nonlinearity: evidence from economic time series", Journal of Econometrics 88:251–281.

Kuo, L., and B. Mallick (1998), "Variable selection for regression models", Sankhya B 60:65–81.

Laird, N.M., and J.H. Ware (1982), "Random-effects models for longitudinal data", Biometrics 38: 963–974.

Lancaster, T. (1997), "Exact structural inference in optimal job-search models", Journal of Business and Economic Statistics 15:165–179.

Leamer, E.E. (1978), Specification Searches: Ad Hoc Inference with Experimental Data (Wiley, New York).

Lenk, P.J. (1999), "Bayesian inference for semiparametric regression using a Fourier representation", Journal of the Royal Statistical Society B 61:863–879.

Li, K. (1998), "Bayesian inference in a simultaneous equation model with limited dependent variables", Journal of Econometrics 85:387–400.

Liu, C., D.B. Rubin and Y.N. Wu (1998), "Parameter expansion to accelerate EM: the PX-EM algorithm", Biometrika 85:755–770.

Liu, J.S. (1994), "The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem", Journal of the American Statistical Association 89:958–966.

Liu, J.S., and R. Chen (1998), "Sequential Monte Carlo methods for dynamic systems", Journal of the American Statistical Association 93:1032–1044.

Liu, J.S., W.H. Wong and A. Kong (1994), "Covariance structure of the Gibbs Sampler with applications to the comparisons of estimators and data augmentation schemes", Biometrika 81:27–40.

Liu, J.S., W.H. Wong and A. Kong (1995), "Covariance structure and convergence rate of the Gibbs sampler with various scans", Journal of the Royal Statistical Society B 57:157–169.

Louis, T.A. (1982), "Finding the observed information matric when using the EM algorithm", Journal of the Royal Statistical Society B 44:226–232.

Mallick, B., and A.E. Gelfand (1994), "Generalized linear models with unknown link function", Biometrika 81:237–246.

Mallick, B., and A.E. Gelfand (1996), "Semiparametric errors-in-variables models: a Bayesian approach", Journal of Statistical Planning and Inference 52:307–321.

Manchanda, P., A. Ansari and S. Gupta (1999), "The "shopping basket": a model for multicategory purchase incidence decisions", Marketing Science 18:95–114.

Marinari, E., and G. Parisi (1992), "Simulated tempering: a new Monte Carlo scheme", Europhysics Letters 19:451–458.

Martin, G. (1999), "US deficit sustainability: a new approach based on multiple endogenous breaks", Journal of Applied Econometrics, in press.

McCulloch, R.E., and R. Tsay (1994), "Statistical analysis of macroeconomic time series via Markov switching models", Journal of Time Series Analysis 15:523–539.

Meng, X.-L., and W.H. Wong (1996), "Simulating ratios of normalizing constants via a simple identity: a theoretical exploration", Statistica Sinica 6:831–860.

Mengersen, K.L., and R.L. Tweedie (1996), "Rates of convergence of the Hastings and Metropolis algorithms", Annals of Statistics 24:101–121.

Metropolis, N., A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller and E. Teller (1953), "Equations of state calculations by fast computing machines", Journal of Chemical Physics 21:1087–1092.

Meyn, S.P., and R.L. Tweedie (1993), Markov chains and stochastic stability (Springer, London).

Meyn, S.P., and R.L. Tweedie (1994), "Computable bounds for convergence rates of Markov chains", Annals of Applied Probability 4:981–1011.

Mira, A., and L. Tierney (1998), "On the use of auxiliary variables in Markov chain Monte Carlo methods", Technical Report (University of Minnesota).

Montgomery, A.L., and P.E. Rossi (1999), "Estimating price elasticities with theory-based priors", Journal of Marketing Research 36:413–423.

Muller, P., and D.R. Insua (1998), "Issues in Bayesian analysis of neural network models", Neural Computation 10:749–770.

Muller, P., A. Erkanli and M. West (1996), "Curve fitting using multivariate normal mixtures", Biometrika 83:63–79.

Nandram, B., and M.-H. Chen (1996), "Accelerating Gibbs sampler convergence in the generalized linear models via a reparameterization", Journal of Statistical Computation and Simulation 54:129–144.

Newton, M.A., and A.E. Raftery (1994), "Approximate Bayesian inference by the weighted likelihood bootstrap (with discussion)", Journal of the Royal Statistical Society B 56:1–48.

Nummelin, E. (1984), General Irreducible Markov Chains and Non-Negative Operators (Cambridge University Press, Cambridge).

O'Hagan, A. (1994), Kendall's Advanced Theory of Statistics, Vol. 2B, Bayesian Inference (Halsted Press, New York).

Paap, R., and H.K. van Dijk (1999), "Bayes estimates of Markov trends in possibly cointegrated series: An application to US consumption and income", Manuscript (RIBES, Erasmus University).

Pastor, L., and R.F. Stambaugh (1999), "Costs of equity capital and model mispricing", Journal of Finance 54:67–121.

Patz, R.J., and B.W. Junker (1999), "A straightforward approach to Markov chain Monte Carlo methods for item response models", Journal of Education and Behavioral Statistics 24:146–178.

Percy, D.F. (1992), "Prediction for seemingly unrelated regressions", Journal of the Royal Statistical Society B 54:243–252.

Pitt, M.K., and N. Shephard (1997), "Analytic convergence rates and parameterization issues for the Gibbs sampler applied to state space models", Journal of Time Series Analysis 20:63–85.

Pitt, M.K., and N. Shephard (1999), "Filtering via simulation: auxiliary particle filters", Journal of the American Statistical Association 94:590–599.

Poirier, D.J. (1995), Intermediate Statistics and Econometrics: A Comparative Approach (MIT Press, Cambridge).

Polson, N.G. (1996), "Convergence of Markov chain Monte Carlo algorithms", in: J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith, eds., Proceedings of the Fifth Valencia International Conference on Bayesian Statistics (Oxford University Press, Oxford) 297–323,.

Propp, J.G., and D.B. Wilson (1996), "Exact sampling with coupled Markov chains and applications to statistical mechanics", Random Structures and Algorithms 9:223–252.

Quinn, K.M., A.D. Martin and A.B. Whitford (1999), "Voter choice in multi-party democracies: a test of competing theories and models", American Journal of Political Science 43:1231–1247.

Raftery, A.E., and S.M. Lewis (1992), "How many iterations in the Gibbs sampler?" in: J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith, eds., Proceedings of the Fourth Valencia International Conference on Bayesian Statistics (Oxford University Press, New York) 763–774.

Raftery, A.E., A.D. Madigan and J.A. Hoeting (1997), "Bayesian model averaging for linear regression models", Journal of the American Statistical Association 92:179–191.

Richardson, S., and P.J. Green (1997), "On Bayesian analysis of mixtures with an unknown number of components (with discussion)", Journal of the Royal Statistical Society B 59:731–792.

Ripley, B. (1987), Stochastic Simulation (Wiley, New York).

Ritter, C., and M.A. Tanner (1992), "Facilitating the Gibbs Sampler: the Gibbs Stopper and the Griddy-Gibbs Sampler", Journal of the American Statistical Association 87:861–868.

Robert, C.P. (1995), "Convergence control methods for Markov chain Monte Carlo algorithms", Statistical Science 10:231–253.

Robert, C.P., and G. Casella (1999), Monte Carlo Statistical Methods (Springer, New York).

Robert, C.P., G. Celeux and J. Diebolt (1993), "Bayesian estimation of hidden Markov models: a stochastic implementation", Statistics and Probability Letters 16:77–83.

Roberts, G.O., and J.S. Rosenthal (1999), "Convergence of slice sampler Markov chains", Journal of the Royal Statistical Society B 61:643–660.

Roberts, G.O., and S.K. Sahu (1997), "Updating schemes, correlation structure, blocking, and parametization for the Gibbs sampler", Journal of the Royal Statististical Society B 59:291–317.

Roberts, G.O., and A.F.M. Smith (1994), "Some simple conditions for the convergence of the Gibbs sampler and Metropolis–Hastings algorithms", Stochastic Processes and its Applications 49:207–216.

Roberts, G.O., and R.L. Tweedie (1996), "Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms", Biometrika 83:95–110.

Rosenthal, J.S. (1995), "Minorization conditions and convergence rates for Markov chain Monte Carlo", Journal of the American Statistical Association 90:558–566.

Rubin, D.B. (1988), "Using the SIR algorithm to simulate posterior distributions", in: J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith, eds., Proceedings of the Fourth Valencia International Conference on Bayesian Statistics (Oxford University Press, New York) 395–402.

Shephard, N. (1994), "Partial non-Gaussian state space", Biometrika 81:115–131.

Shephard, N. (1996), "Statistical aspects of ARCH and stochastic volatility", in: D.R. Cox, D.V. Hinkley and O.E. Barndorff-Nielson, eds., Time Series Models with Econometric, Finance and Other Applications (Chapman and Hall, London) 1–67.

Shively, T.S., R. Kohn and S. Wood (1999), "Variable selection and function estimation in additive nonparametric regression using a data-based prior", Journal of the American Statistical Association 94:777–794.

Smith, A. (1999), "Testing theories of strategic choice: the example of crisis escalation", American Journal of Political Science 43:1254–1283.

Smith, A.F.M., and G.O. Roberts (1993), "Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods", Journal of the Royal Statistical Society B 55:3–24.

Smith, M., and R. Kohn (1996), "Nonparametric regression using Bayesian variable selection", Journal of Econometrics 75:317–343.

So, M.K.P., K. Lam and W.K. Li (1998), "A stochastic volatility model with Markov switching", Journal of Business and Economic Statistics 16:244–253.

Stephens, D.A. (1994), "Bayesian retrospective multiple-changepoint identification", Applied Statistics 43:159–178.

Tanner, M.A. (1996), Tools for Statistical Inference, 3rd. edition (Springer, New York).

Tanner, M.A., and W.H. Wong (1987), "The calculation of posterior distributions by data augmentation", Journal of the American Statistical Association 82:528–549.

Taylor, S.J. (1994), "Modelling stochastic volatility", Mathematical Finance 4:183–204.

Tierney, L. (1994), "Markov chains for exploring posterior distributions (with discussion)", Annals of Statistics 22:1701–1762.

Tierney, L., and J. Kadane (1986), "Accurate approximations for posterior moments and marginal densities", Journal of the American Statistical Association 81:82–86.

Tsionas, E.G. (1999), "Monte Carlo inference in econometric models with symmetric stable disturbances", Journal of Econometrics 88:365–401.

Verdinelli, I., and L. Wasserman (1995), "Computing Bayes factors using a generalization of the Savge–Dickey density ratio", Journal of the American Statistical Association 90:614–618.

Wakefield, J.C., A.F.M. Smith, A. Racine-Poon and A.E. Gelfand (1994), "Bayesian analysis of linear and non-linear population models by using the Gibbs sampler", Applied Statistics 43:201–221.

Waller, L.A., B.P. Carlin, H. Xia and A.E. Gelfand (1997), "Hierarchical spatio-temporal mapping of disease rates", Journal of the American Statistical Association 92:607–617.

Wei, G.C.G., and M.A. Tanner (1990a), "Posterior computations for censored regression data", Journal of the American Statistical Association 85:829–839.

Wei, G.C.G., and M.A. Tanner (1990b), "A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithm", Journal of the American Statistical Association 85:699–704.

West, M., R. Prado and A.D. Krystal (1999), "Evaluation and comparison of EEG traces: latent structure in nonstationary time series", Journal of the American Statistical Association 94:375–387.

Wu, C.F.J. (1983), "On the convergence properties of the EM algorithm", Annals of Statistics 11:95–103.

Young, M.R., W.S. DeSarbo and V.G. Morwitz (1998), "The stochastic modeling of purchase intentions and behavior", Management Science 44:188–202.

Zeger, S.L., and M.R. Karim (1991), "Generalized linear models with random effects: a Gibbs sampling approach", Journal of the American Statistical Association 86:79–86.

Zellner, A. (1971), Introduction to Bayesian Inference in Econometrics (Wiley, New York).

Zellner, A., and C. Min (1995), "Gibbs sampler convergence criteria", Journal of the American Statistical Association 90:921–927.

This Page Intentionally Left Blank

Part 13

APPLIED ECONOMETRICS

This Page Intentionally Left Blank

Chapter 58

CALIBRATION

CHRISTINA DAWKINS [*]

University of Warwick

T.N. SRINIVASAN [**]

Yale University

JOHN WHALLEY [*]

Universities of Warwick and Western Ontario and NBER

Contents

Abstract 3654
Keywords 3654
1. Introduction 3655
2. Calibration: its meaning and some early examples 3656
2.1. Two examples 3657
2.2. Reasons for using calibration 3660
3. The debate about calibration 3661
3.1. Lines of the debate 3662
3.2. Calibration is estimation, estimation is calibration 3663
3.3. The Jorgenson–McKitrick critique of micro model calibration 3665
4. Making calibration more concrete 3667
4.1. Calibrated macroeconomic models 3667
4.2. Calibrated microeconomic models 3671
4.3. Deficiencies of micro calibration 3677
4.4. Some recent examples of calibrations 3677
5. Best practice in calibration 3684
5.1. The choice of model for calibration 3684

[*] The first and third authors acknowledge financial support from the UK Economic and Social Research Council under an award for General Equilibrium Analysis of UK Policy Issues.

[**] We thank James Heckman, Edward Leamer, John Rust and Gianluca Violante for their helpful comments. We are also grateful to Andres Erosa and Gustavo Ventura for comments on an earlier draft.

[*] The first and third authors acknowledge financial support from the UK Economic and Social Research Council under an award for General Equilibrium Analysis of UK Policy Issues.

Handbook of Econometrics, Volume 5, Edited by J.J. Heckman and E. Leamer

5.2. The calibration of what to what? 3686
5.3. The choice of functional form 3687
5.4. The use of elasticity parameters 3689
6. New directions in calibration 3694
6.1. Expost decomposition and double or multiple calibration 3694
6.2. Sensitivity analysis 3695
6.3. Preadjusting data 3697
7. Conclusion 3698
References 3699

Abstract

We discuss the use of calibration techniques in economic models. Calibration contrasts with estimation in relying on deterministic calculation of model parameter values consistent with data, rather than econometric estimation. The reasons why calibrators use these methods, as well as the main arguments in debates between calibrators and econometricians are set out. We draw a distinction between the calibration methods used in dynamic macro models of the Kydland–Prescott type and micro models of the Shoven–Whalley variety. We highlight the ways in which calibration techniques are evolving including double calibration, the use of data pre-adjustments, and the incorporation of model estimation consistent elasticities. We conclude with a discussion of what constitutes best practice in calibration.

Keywords

benchmarking, best practice methods, calibration, dynamic macromodels, econometric estimation, elasticity parameters, equilibrium data set, real business cycle models

JEL classification: C13, C19, C68, C80, C81, C82

1. Introduction

Thirty years ago, the suggestion that non-econometrically estimated numerical models with parameter values taken partly from other studies and the remainder determined so as to be consistent with observed data, could pass as serious empirical work in economics would have been greeted with a certain amount of incredulity. Econometric theory seemed to be advancing towards ever more completeness and, despite the continued theoretical focus on deterministic modelling, a statistical component to any economic model applied to data was assumed to be indispensable.

But in the late 1990s, despite the substantial controversy over their use, so-called "calibrated" models, which are best described as numerical models without a complete and consistent econometric formulation, have become a central element of empirical work in economics. This growth has occurred even though the term "calibration" has been used to denote a variety of procedures. First, in applying general equilibrium microeconomic models to policy evaluation, and later in a variety of other areas including stochastic general equilibrium non-monetary macroeconomic models seeking to investigate the causes and consequences of business cycles, models incorporating some form of calibration procedure have spread through the literature. One survey of applied work in macroeconomics, Gregory and Smith (1991), even declared such models to be the predominant tool in contemporary macroeconomics for empirical investigation.

The use of calibration in economics has generated deep controversy in the profession about the relative merits of calibrated and estimated models. While we offer a brief summary of the accompanying debate, our main objective here is to move beyond it. In our view, calibrated models are here to stay. We ask what is calibration, and how does it relate to econometric estimation and/or testing. We note that while many authors describe their procedures for determining parameter values as calibration, few are explicit about what calibration actually involves. What comprises best practice in calibration? What is to be calibrated to what? Is calibration really as radically different from econometric procedure as has often been asserted in the literature? Should the use of the word calibration be restricted to procedures which ensure that a complete base case or benchmark equilibrium data set is always reproduced as an equilibrium of the calibrated model, or can it also be extended to situations where parameter values in models are set in a more arbitrary manner? Does the best form of calibration also depend on the question to be explored in the research? What are the weakest links in calibration procedures, and what can be said about the notoriously imprecise use in calibrated models of elasticities drawn from the literature? Finally, what are some of the new directions in calibration?

Section 2 of this chapter explores the meaning of the term, and presents some early examples of calibration. Section 3 outlines the debate surrounding calibrated models. Section 4 provides concrete illustrations and discusses how calibration procedures are implemented in more recent work. Section 5 highlights current issues in calibration, and Section 6 sets out some of the new calibration procedures being used. Section 7 concludes.

2. Calibration: its meaning and some early examples

The term calibration denotes the setting of the origin and choice of scale for a measuring instrument; a thermometer calibrated to read 0 C and 100 C when water freezes and boils, can be used to measure temperature. Calibration of an economic model involves the setting of specified parameters to replicate a benchmark data set as a model solution[1]. Once calibrated, the model can be used to assess the effects of an unobservable or counterfactual change in policies or other parameters: a change in a tax rate, the removal of a quota, or changes in the values of parameters exogenous to the model such as the prices of traded goods in a model of a small, open economy. The model's counterfactual solution provides a measure of what the change may produce. It offers a prediction of the way in which the economy is likely to respond to the change, while the model's base case or pre-change solution is the reference point – the observed outcome from the economy under the existing policy regime and values of the exogenous parameters.

The driving force behind the use of calibration in economics is the belief that any counterfactual analysis is impossible without a coherent theoretical framework and that models which are consistent with economic theory are the place to start. In empirical research one often tests how well a particular model describes the data or, more generally, one selects the model that best describes the data. But models that are estimable as a single entity and permit testing or selection are usually relatively simple. Policy analysis often requires the specification of more complex models which preclude estimation or testing. Such models, viewed as "theory with numbers", involve the empirical task of parameterizing rather than testing models. If parameters are not available in the literature, and if the model as a single entity is unestimable, parameter values must still be obtained somehow. The term calibration generally indicates the use of procedures that implement this parameterization requirement.

Calibration, however, remains an imprecise term despite its widespread use. No single set of calibration procedures exists, nor does the term indicate what is being calibrated to what. Thus, micro general equilibrium modellers frequently calibrate models to a single (constructed) equilibrium observation. The idea is to generate a model specification capable of reproducing the constructed data as a model solution (a computed general equilibrium). However, basic data, typically drawn from several

[1] In calibrating a general equilibrium model, for example, the numerical values of some model parameters are typically set exogenously, while others, the calibrated parameters, are endogenously determined so as to reproduce the benchmark data as an equilibrium of the model. The exogenously specified parameters are typically the elasticities of substitution in constant elasticity of substitution (CES) functional forms, which are usually set on the basis of estimates drawn from the literature.

sources, will not satisfy the equilibrium conditions of the model[2], and adjustments have to be made in the basic data prior to calibration to construct a microconsistent equilibrium data set. This adjustment procedure is usually called "benchmarking" and the adjusted data set, the "benchmark" data set. In addition to generating model specifications, this type of calibration exercise also allows the benchmarked data to be used as a consistency check for the model solution procedures.

Highly aggregated business cycle models specify structures that include stochastic elements of the model which influence behaviour. Their structure typically includes a steady-state or long run joint distribution of the aggregate variables of the model that is described parametrically. Calibration in this context, consists of asking whether, for plausible values of its parameters, the steady-state distribution generated by the model corresponds to that of the data.

To calibrators in economics, econometric procedures often appear to be based on theoretically poorly specified models of behaviour and seldom appear to produce conclusive results. Policy and other issues of the day cannot wait until the theoretical models needed to analyze them are well developed in the literature. All that is needed is to specify an appropriate model, and choose values for its parameters. Calibration is a process that produces this outcome, even if it sometimes yields parameter values which are not derived from time series or cross section estimation.

2.1. Two examples

Two well known early examples in the literature illustrate the rationale for and procedures used in calibration. In one of the earliest calibration exercises, Shoven and Whalley (1972), were attempting to refine Harberger's earlier (1962) calculations of the welfare cost of differential tax treatment of capital income by sector in the US. They used Harberger's earlier model and data (averaged over years in the late 1950s and early 1960s, with 1959 as the mean), but applied Scarf's (1973) algorithm to solve the model for exact equilibria rather than the approximate equilibria that Harberger had obtained by linearizing around an initial pre-tax change equilibrium. They took Harberger's data and extended it via a few simple adjustments into a benchmark equilibrium data set.

Initially they tried to generate numerical values for the parameters of their model by adjusting initial values iteratively, and seeing how closely the model solutions reflected the constructed equilibrium data set. Early working paper versions presented diagrams which illustrated the distance between the data and the model solutions arising from the use of various combinations of parameters. Shoven and Whalley then realized that instead of simply trying alternative combinations of parameters, they could use the equations characterizing an equilibrium solution of the model to solve for the values of the parameters whose values had not been set exogenously. In essence, their

[2] Equilibrium conditions include, for example, demand-supply equalities and zero-profit conditions, whenever constant returns to scale and pure competition prevail.

procedure converted these parameters into variables, and solved for their values by trivially imposing equilibrium as an identifying restriction, using benchmark data.

One reason why Shoven and Whalley adopted this procedure in their work was that other fledgling numerical general equilibrium models of the time had relied exclusively on literature based estimates for all parameter values, and importantly had used them in the base case specification of the model. Thus, in an economy contemplating tax reform, such a model specification might give a base case solution in which, for example, 50% of employment was in manufacturing, when the data clearly showed the figure to be 25%. Thus, the actual performance of the economy in the base case, which was known from national accounts data, was in no way reflected in the base case solution of the model with all parameter values taken from the literature. Their observation that model outcomes differed from data led them to reject the exclusive use of literature based values for all parameters and to allow, instead, the values of a subset of parameters to be generated by the model structure and the requirement that the benchmark data represent an equilibrium solution of the model. Thus, in their calibration procedure they used the equilibrium solution of their model as an identifying restriction to obtain numerical values for a subset of parameters. Their calibration had no predictive power since a variety of models, functional forms, and alternative subsets of free parameters could be calibrated to the same data. However, the first requirement of the model was to reproduce exactly or closely the known base data as an equilibrium in what has become known as the replication test[3]. Failure to do so, caught many coding or other errors. This approach is reflected in the natural sciences, where the performance of equipment is tested by having it replicate the known solution to a problem.

A second early example of calibration is that of Kydland and Prescott (1982) in their first real business cycle model. It is a simple one sector growth model with labour–leisure choice and non-time separable preferences, which they argued could be used to explain the autovariances of real output and the covariances of cyclical output with other aggregate time series for the post-war US economy. The crucial element of their structure, given their purpose, was the assumption that more than one time period is needed to construct newly productive capital. They contrasted this treatment with the then more conventional aggregate investment functions based on an assumed adjustment cost as a function of the value of investment, criticizing the adjustment cost approach on the grounds that the time required to complete investment projects is not short relative to the business cycle. Labour supply and new investment decisions at time t were thus contingent on the past history of productivity shocks, the capital stock at time t, and the marginal utility of leisure parameter.

Kydland and Prescott then introduced stochastic technology shocks into their structure, through a Hicks neutral shock to the aggregate production function which consisted of a permanent and a transitory component. The permanent component

[3] The replication test is undertaken assuming the absence of multiple equilibria.

was, by definition, highly persistent; and so the transitory component was equal to the transitory shock. A third shock was a disturbance to productivity. This generated a recursive informational structure, and the general structure followed a vector autoregressive process with independent normal innovations.

Kydland and Prescott argued that a test of their structure was whether a set of parameters existed for which the model's co-movements for both the smoothed series and the deviations from the smoothed series were quantitatively consistent with the observed behaviour of the corresponding series for the post-war US economy. They added the further requirement that the parameters chosen should not be inconsistent with the relevant micro observations, including the reported construction periods for new plants and cross-sectional observations on consumption and labour supply. They suggested that the closeness of their specification of preferences and technology to those used in related applied work facilitated comparisons to other work.

They first specified their model so that its steady state properties were consistent with long term trend data for the US. Quantitatively explaining the co-movements of the deviations from trend remained as the test of the underlying theory. They emphasized some of these key co-movements; investment varied three times as much as output while for consumption the variation was only one half; variations in output largely reflected variations in hours worked per household, not capital stocks or labour productivity.

Kydland and Prescott provide a three page discussion of how they calibrated their model, by choosing the majority of their parameter values and leaving other parameters free to be determined by a model fit to data. The first bloc of parameters was largely chosen by appealing to plausible values for key aggregates and literature estimates. Two parameters affecting the intertemporal substitutability of leisure and three variance parameters on productivity shocks were left free, with the sum of the variance parameters restricted so that the model estimate of the variance of cyclical output equalled that of the US economy.

For each set of parameter values, the autocorrelation of cyclical output for up to six periods was computed, along with standard deviations of cyclical variables of interest and their correlations with cyclical output. These were compared to the same statistics for the US economy. Kydland and Prescott chose what they considered to be the best fit and then examined the actual model solutions. Comparing estimated autocorrelations for real output from the model with sample values for the US economy, Kydland and Prescott concluded that the fit was surprisingly good. On this basis, they suggested that the model loosely met a goodness of fit criterion, and could be accepted as a reasonable structure to use to analyze macro issues in the US. Put another way, in their model closeness to observed values of a specified set of autocorrelations and correlations was the identifying restriction, just as the requirement that the model solution reproduces the benchmark data as an equilibrium served as the identifying restriction in the Shoven–Whalley models of Walrasian general equilibrium.

These two early examples serve both to illustrate the calibration approach and to highlight the diversity in its application and the inferences drawn. Subsequent calibrators using micro policy evaluation models, like Shoven and Whalley, frequently construct microconsistent data sets which are fully compatible with the equilibrium conditions of their model, and then calibrate their models to these exactly. They choose values for parameters of preferences and technology in their models, so that once all parameter values, specified and as well as calibrated, are entered into the model, solving the model yields an equilibrium solution which is identical to the benchmark microconsistent data. This "replication check" is analogous to testing an algorithm for problem solving by checking that it can solve a problem to which the answer is already known. A failed replication check can signal coding or other errors. Thus, the microconsistent data sets are used twice: the equilibrium conditions embedded in them are first used for the calibration exercise of solving for the numerical values of free parameters; then for the second time in the replication exercise to verify that there are no errors either in the calibration exercise or in the model solution algorithm and procedure.

Calibrators using dynamic macro models typically take a different tack. They check the value of model parameters for their ability to generate equilibrium stochastic time paths for steady-states (as well as transitions) that are consistent with the stochastic properties of the joint distribution of the observed data on the same aggregate. Thus, they evaluate their model structure on the basis of how closely the model solution approximates real data. Observed data are not, and indeed cannot, be used to infer parameter values which exactly replicate base data, nor are equations characterizing the model solution used to solve for model parameters with the role of endogenous and exogenous parameters reversed. Furthermore, unlike the micro policy evaluators, they make no preadjustments to their basic data, assuming in effect, that the data represent realizations of the equilibrium path of some model with the same structure as the one they use. This type of calibration has its origins in the real business cycle literature which followed Kydland and Prescott's contribution, where the closeness of particular moments in the model solutions and the data is the key. Unlike in micro policy analyses, where no stochastic disturbances are admitted, an important feature of macro modelling is to recognize that the economy is subject to random shocks so that observed data are stochastic. Hence, they look for model parameter values that generate stochastic distributions matching those implied by the data.

2.2. Reasons for using calibration

The factors which have caused researchers to adopt these and other calibration methods are many, and as can be seen from the two cases discussed above, vary substantially from case to case. One factor has been the growing interest of applied economists in models with richer structures than are currently found in many econometric models. The unsatisfactory state of play with macro econometric models, and the desire to have models which are consistent with observed co-variation between output, consumption and investment, were a key factor for Kydland and Prescott. Finding little in the conventional literature that suited their purpose, they developed a new approach.

Equally, Shoven and Whalley, finding no estimated general equilibrium models which could be used for policy investigation, also developed their own approach.

Calibration has come into play because the economics in econometrics, and the economics in pure theory seem to have progressively drifted apart. A description, or perhaps a caricature of some of the recent applied econometric work would be as follows: an elaborate non-linear deterministic theoretical model is first discussed, and then linearized (often without clearly specifying around what point the model is linearized), stochastic disturbances are added before proceeding to estimation with data. Substantial econometric sophistication is brought to bear on the specification of the stochastic properties of the disturbances, which are in reality only *ad hoc* additions to the basic model. The underlying economics in such econometric work is often so simple that without the statistical component, the model would be rapidly discarded. Yet the momentum to further these econometric refinements continues. Thus, demand estimation has advanced from single commodity demand functions to systems of demand functions; but combined demand and supply systems are rarely estimated, multi-consumer demand systems seemingly not at all, and the two person pure exchange economy remains without any econometric application. Econometric models of demand, until recently, typically did not incorporate such features as product quality, product characteristics and other features – many of which are issues that numerical modellers feel they have to incorporate into their models.

Calibrated models have long been employed in disciplines other than economics, such as physics, resulting in an interplay between theoretical developments and attempts to assess whether new theoretical structures account for actual observations by applying the latest best guesses for parameter values to theoretical structures. Their growth in use in economics to some degree reflects this evolution in other disciplines and should not be seen as an aberration, but rather as an empirically focussed scientific investigation parallel to and supportive of, if different from, econometric investigation.

3. The debate about calibration

As the calibration literature has developed, the use of calibration has been accompanied by substantial controversy. Why are calibrated models not estimated and subjected to econometric testing, like other empirically based models in economics? How much reliance can be placed on the use of literature based parameters since these are frequently unavailable, and where they exist have wide ranges, or are often contradictory? Is not calibration clearly inferior to estimation?

This controversy has resulted in a series of recent papers and symposia in major journals in most cases centering on the use of calibration in macro models. A symposium on calibration in the 1996 *Journal of Economic Perspectives* saw Kydland and Prescott set out their interpretation of calibration, with commentary from Hansen and Heckman, and Sims. Symposia in the 1995 *Economic Journal,* edited by

Quah, and the 1994 *Journal of Applied Econometrics,* edited by Pagan, are further contributions. Some papers with provocative titles (such as Hoover's (1995) "Facts and Artifacts" paper), have appeared. De Jong, Ingram and Whiteman (1996), who discuss a Bayesian approach to calibration, indicate new approaches. To Gregory and Smith (1991), the issue is calibration as estimation.

3.1. Lines of the debate

Hoover (1995) sets out some of the lines of the debate on calibration; discussing the empirical value of calibrated models over estimation. The criticisms of calibration range from claims of casual empiricism – that parameters from unrelated econometric studies are used and the fact that models are not formally tested – to assertions that reduced form methods deliver more empirically. Watson (1993), for instance, emphasizes that in macro calibrations the metric used to determine the distance between the simulated and actual moments in data is often left unspecified[4]. Sims (1996) explains why, in his view, dynamic stochastic general equilibrium modelling has delivered little by way of empirical payoff; macroeconomists have developed a variety of other approaches to compressing time series data using only informal theoretical approaches[5].

Hoover cites Lucas (1987) as providing key counter arguments. Lucas argues that the question of whether a model is true is not particularly interesting, because all models are clearly abstractions. They are meant to be workable tools to answer a limited set of questions. But a model which is not true can still be used to derive useful quantitative guides to policy. Consequently for Lucas, and for Kydland and Prescott, testing a model is uninformative. This position contrasts with that of Sargent as a representative of the school of estimation, which holds that testing is necessary to evaluate alternative models.

Hoover, however, presents two criticisms of the practice of calibration in Kydland and Prescott. The first is that of "casual empiricism". In contrast to the calibration of micro models described in Mansur and Whalley (1984), which relies on systematic literature searches to find values for elasticity parameters, Kydland and Prescott's approach to choosing selected parameters seems casual to Hoover. Other macro modellers have acknowledged this criticism. The recent discussion in Browning et al. (1999) explores and responds more fully to this concern. The second criticism is that Kydland and Prescott do not provide any formal measure of the performance of their model, although this issue has been addressed in recent literature such as Watson (1993), who presents a measure of fit for calibrated models.

[4] This point does not apply to micro calibrators, because they calibrate their models to data exactly.

[5] But again, in the micro areas, numerous insights on the significance of policy measures have been generated by calibrated models. These range from the efficiency and distributional impacts of tax policies, to the impacts of trade policies in various economies.

The debate between estimators and calibrators has also revealed other key differences in approach. Estimators, Hoover argues, pursue a competitive strategy in which "theories compete with one another for the support of data" [Hoover (1995, p. 29)]. The rejection of a model on statistical grounds casts doubt on the validity of its underpinning theory. In contrast, calibrators use an adaptive strategy which has at its heart a parsimonious, idealized model derived from theory. The calibrator extracts all possible information from that model. Where simulation outcomes do not match data to the extent desired by the modeller, the underlying theory is not rejected, but instead the modeller adds features in an attempt to improve the match. The merits of adding specific features is gauged by the subsequent improvements in the model's performance in such matching. Hoover notes that unlike econometric models, calibrated models cannot shed light on which of two fundamentally different models is best, because they are not subjected to formal testing.

Despite the growth in its use, Hoover argues that no compelling defence has been made for the calibration methodology. He turns to the concept of a model advocated by Lucas for the elements of such a defence: to be comprehensible and useful, a model must be simple, and hence abstract; but on the other hand it must capture sufficient features of reality to inspire confidence in its ability to shed light on empirical questions. Underlying the econometric approach, with its concern about issues such as omitted variable bias and specification error, is the assumption that any empirical methodology which does not strive towards fully articulated models, will generate misleading results. However, the answers to some questions, such as welfare or efficiency changes, are not always easily subjected to statistical tests. In these cases, simple models which mimic salient features of the real world can offer otherwise unavailable quantitative insights.

3.2. Calibration is estimation, estimation is calibration

Despite their portrayal in the literature our position is that calibration and estimation are in less conflict than one might suppose. If calibration is the setting of the numerical values of model parameters relative to the criterion of an ability to replicate a base case data set as a model solution, and estimation is the use of a goodness of fit criterion in the selection of numerical values of model parameters, the two procedures are closely related. In both cases a selection of model parameter values which is thought to be reasonable (or best) relative to some criterion applied to data is involved. In one sense, both procedures lead to identical outcomes.

Suppose in an econometric estimation exercise, say estimating a linear regression, there are more explanatory variables, and hence more regression parameters, than the number of observations. In general, in this situation of underidentification, there will be more than one set regression coefficients that will explain the data exactly and lead to an R^2 of 1. The situation of parameter calibration applied to general equilibrium models is analogous, since the models typically have more parameters than data and, as such, an exact fit is possible. Furthermore, more than one set of parameter values will typically fit the data exactly.

Where the differences lie is in the type of models used and the criteria applied in the process of fitting models. Estimation is applied to models that often have a limited economic structure, but for which the statistical extension can be complex. The criteria used in estimation are many and, like the maximization of likelihood functions, can be highly sophisticated. These structures allow practitioners to undertake a variety of statistical tests about the properties of the estimated parameters and the performance of the model in light of the data. All statistical tests, however, are necessarily conditional on some maintained, but untested, hypotheses.

Calibrators also use their models for different purposes than econometricians. Whalley (1985b) suggests that micro general equilibrium models may well be impossible to estimate in conventional terms, except perhaps in a situation where an economy with unchanged parameters is observed in equilibrium several times as observed exogenous variables change. Even the basic two-person pure exchange economy remains unestimated, because the cross-equation restrictions implied by the central model solution concept of equilibrium are too complex to impose on conventional estimation. A pure exchange economy requires excess demands which satisfy the conditions for the Brouwer fixed point theorem; an economy with production requires supply correspondences and demand functions which satisfy the conditions for the Kakutani fixed point theorem. Any estimation of a model of a general equilibrium economy must generate parameter estimates that satisfy market clearing, zero profit conditions, and Walras Law. Thus the growth of calibration, in part, reflects the relative lack of fully estimated models for use in applications of the theory held by theorists to be so central to policy and other analyses.

Furthermore, the implications of viewing the data as representative of a behavioural equilibrium or some well specified disequilibrium dynamic process are typically not imposed as constraints in estimation. In many econometric models the stochastic error terms are *ad hoc* additions, which are not formally part of the decision making of the agents which the models purport to represent. Like Lucas, Prescott (1986) rejects the notion of testing econometric models as unhelpful on the basis that any economic model is a simplified abstraction from a more complex reality, and hence any model can be nested within a more complex, but more realistic, variant. Testing models against each other is not the issue; the question is which simplified model to use; which best captures the essence of the economic processes one wants to analyze. Judgement based on the model's use, is an essential ingredient in this choice.

What is implicitly advocated in both of these cases is a different approach instead of the replacement of estimation by calibration. The objective is the same, the best fit of a model to data; the difference lies in the criteria and the model structures, and hence the procedures employed. Because in the case of Whalley, the unestimable or unestimated nature of models central to economic theory is taken as the starting point, and in the case of Prescott, model testing is rejected, both appeal to other criteria for model selection. These criteria do not rely on econometric or statistical techniques. Whalley argues for selecting models that are widely used in the theoretical literature but have no econometric analogue. Prescott advocates selecting models which are theoretically well grounded, and standard in the theoretical literature.

The choice of model structure, in the absence of either econometric estimation or meaningful tests between models, is therefore based on appeals to widespread use in theoretical work. This, as much as anything else, is the growing divide between calibration and applied econometrics. The purpose of empirical investigation using models is less to test them, than it is to assess their numerical implications. Model structures, in turn, need a numerical specification; leading directly to calibration.

3.3. The Jorgenson–McKitrick critique of micro model calibration

Despite the interrelationships between calibration and estimation we discuss above, micro calibration, in particular, has nonetheless been criticized in Jorgenson (1984) and more recently by McKitrick (1995) on several counts. One is that the data pre-adjustments in the process of implementing calibration introduce untraceable bias into the data and hence, into the model results. The use of a benchmark year for calibration also enters their critique. Because calibration relies on a single observation, any anomalies in the economy for that year can be transmitted to the model results, and may taint the conclusions. They highlight the inadequacies of the elasticity estimates in applied models, and argue that the reliance on CES, Leontief, and Cobb–Douglas functional forms is restrictive and unrealistic. This restrictive class of functional forms precludes complementarities, incorporates elasticities of substitution that are independent of prices and which, thus, unrealistically constrain behavioural responses in counterfactual simulations.

In response to these shortcomings, they suggest the simultaneous estimation of all of a model's elasticities and share parameters using time series data. This estimation of general equilibrium parameters is, however, largely undertaken for model subsystems, rather than by incorporating the full set of cross-equation equilibrium restrictions into the procedure. Such an explicit econometric approach to general equilibrium modelling has thus far been limited to a handful of papers: Clements (1980), Jorgenson (1984), Jorgenson, Slesnick and Wilcoxen (1992), and McKitrick (1995). This approach, they suggest, allows for elasticity estimation which is fully consistent with the definitions of variables employed in the model and does not require the use of restrictive functional forms. The statistical basis of their estimation isolates systematic effects from random noise and the use of unadjusted time series data precludes the introduction of pre-adjustment bias.

If this approach is superior to calibration, the question which arises is why has it not been more widely adopted. One issue is the estimability of equilibrium models, raised above. How is the equilibrium solution concept central to general equilibrium analysis actually going to be imposed as a series of cross-equation restrictions in complete model estimation? A second involves the use of statistical rather than deterministic models. If a stochastic growth model of the form used in the real business cycle literature is calibrated to data, but not tested, the calibration in applying goodness of fit criterion is really only a variant of estimation, and vice versa. Other reasons for the limited use of econometric methods in the parameterization of general equilibrium models lie in the paucity of time series data on the variables of interest for the questions that are addressed in calibrated models. Estimating large dimensional models, or models which focus on variables that are not measured in national accounts data, may be intractable. The effort required to generate the single observation required for calibration can itself be formidable, and extending the process to include time series observations may be close to impossible. For example, modellers must frequently update an earlier year's input–output matrix as an approximation to that of the benchmark year, because in many economies annual input–output tables are not produced. Furthermore, where they are produced, they are often generated by updating a previous year's table rather than by undertaking new production surveys.

The econometric approach also precludes the use of some simplifying techniques commonly employed in applying general equilibrium models. One such technique is the Harberger (1962) convention, whereby the units of quantities defined in the model are given by that quantity which sells for one unit of currency in base period data. This convention allows the modeller the simplification of representing heterogenous quantities in a homogenous manner, both in data and in the model. For example, if labour inputs were to be measured as hours worked, some correction would have to be made for different levels of labour efficiency and skill. The use of this assumption also reduces the number of variables required in the model; the modeller need only collect data in value terms, rather than in separate price and quantity terms. Such a convention, however, creates time-dependent units and makes time series estimation all but impossible.

Modellers have responded to certain elements of the Jorgenson–McKitrick critique, within their calibration paradigm. The weakness of the elasticity estimates has been addressed via the sensitivity analysis procedures in Wigle (1991), Pagan and Shannon (1985, 1987), DeVuyst and Preckel (1997), Harrison and Vinod (1992), and Harrison et al. (1992), discussed in Section 6 below. Modellers need no longer rely on restrictive functional forms; a fully flexible, globally regular functional form, has been developed by Perroni and Rutherford (1998).

The issues of the adjustments made to data for calibration purposes have been largely ignored in the modelling literature, but the pitfalls of drawing conclusions from a single and possibly unrepresentative, single year benchmark observation have been explored in several papers. Roberts (1994) examines the significance of this in a model of Poland by calibrating to five different benchmark years, and concludes that model results are robust to the choice of benchmark year. Adams and Higgs (1990) also address this problem, arguing that the effects can be mitigated by averaging several years' data into a single observation. They also illustrate how using agricultural data from an abnormal "year of record" affects policy conclusions derived from the Australian ORANI model. Dawkins (1997) develops a methodology for undertaking sensitivity analysis with respect to the initial data values.

The introduction of untraceable bias to the model parameterization through preadjustments remains a largely unaddressed issue. The only example to our knowledge is Wiese (1995), who derives two benchmark equilibrium data sets using alternative accounting assumptions for employer contributions to health insurance and traces the effects of these assumptions on model results. His experiments indicate that the model results are indeed affected by the accounting conventions used in the data. Different accounting conventions could, in principle, also affect econometric estimates in so far as such conventions serve as identifying restrictions. The bias introduced into model results through the use of algorithms employed in the systematic component of data adjustments has been recently explored in Dawkins (1998).

Through calibration, any adjustment of data potentially has implications for the model's results, but these biases may be virtually indiscernible where adjustments are performed on an *ad hoc* basis. While escaping the piecemeal approach to data preadjustment entirely may be impractical, a move towards more systematic and better reported data adjustment is desirable so that ways of addressing these bias issues can be developed within the calibration framework.

4. Making calibration more concrete

4.1. Calibrated macroeconomic models

Calibration entered the macroeconomics literature with Kydland and Prescott's 1982 paper. Apart from its technical dexterity, the extraordinary originality in both approach and execution in this paper continues to galvanize interest more than a decade and a half later. Few other papers over the last fifty years have made such a strong impact, drawn so much admiration, and from other quarters, perhaps had so much misinformed adverse comment targeted their way. Although much of the subsequent debate over Kydland and Prescott's paper has been about calibration, calibration was not their main focus. The word itself does not appear until the latter half of the paper, and then it only appears once, in the subheading "Model Calibration". Nowhere does the word appear in the text of the published paper, and no explanation of the term is offered.

As discussed above, Kydland and Prescott sought to show how a stochastic growth model could be developed and fitted to post-war US quarterly data in which the comovements of the fitted model are quantitatively consistent with the corresponding comovements found in US data. Their objective was to show that an essential component of explaining aggregate fluctuations is the recognition that producing and installing capital equipment and bringing it into use cannot occur in the same period as the original decision to invest. Kydland and Prescott noted that macro-econometric models offered them little help with model parameterization; wholesale estimation of the model appeared infeasible and choosing parameters as they did, seemed to be a reasonable approach for the problem at hand.

The technique they used was to construct a model with technology and non-time separable preferences, setting out the time required to build new productive capital, and an aggregate production function incorporating technology shocks. They characterized the steady state of their system, and showed how its general structure was akin to a vector autoregressive process with independent, multivariate normal innovations. Steady state behaviour of the model was tied down by the specification of aggregate functions, and its stochastic structure was chosen to fit as closely as possible to the actual covariation in the data. As noted earlier, given their specification of preferences, production structure and stochastic structure, Kydland and Prescott concluded that the fit of model results to actual data was surprisingly good.

The Kydland and Prescott calibration set off debate on an ever-widening set of questions concerning the calibration of macro models. Which models should be considered as admissible for the purposes of such an investigation, and how should these be calibrated? Should calibration be restricted to the long run growth path of the economy, and how should it be implemented? Only limited reference was made to the earlier exact calibration procedures used by the micro-based policy modellers. In the original Kydland and Prescott scheme, for instance, if stochastic shocks were to apply to a wider range of parameters, such as Cobb–Douglas shares in production and demand parameters, with sufficient freedom over the specification of the stochastic processes involved, exact calibration to the distribution of observed aggregate data could likely have been performed.

Cooley and Prescott (1995) have recently set out in more detail what such calibrations imply and have suggested that since the underlying structure used in real business cycle models is the neoclassical growth framework, choosing parameters and functional forms through calibration is only designed to ensure that the underlying model economy will display steady-state or balanced growth behaviour that is consistent with actual data. They suggest preserving this calibration standard in studies of business cycles, although they stress that this standard does not imply that the model economy, in its stochastic form, will reproduce actual business cycle behaviour.

As an example, they present an economy with a single, infinitely lived, dynastic consumer whose size grows at the rate η; where γ is the long term growth rate of labour-augmenting technical change. Preferences and technology are both assumed to be Cobb–Douglas. The dynastic consumer maximizes expected utility from consumption, and forms expectations on future prices to enable it to do so. The dynastic optimization problem is set out by Cooley and Prescott as

$$\max E\left[\sum_{t=0}^{\infty}\beta^t(1+\eta)^t[(1-\alpha)\log c_t+\alpha\log(1-h_t)]\right], \tag{1}$$

$$\text{s.t. } c_t+x_t=e^{z_t}(1-\gamma)^{t(1-\theta)}k_t^{\theta}h_t^{(1-\theta)}, \tag{2}$$

$$(1+\gamma)(1+\eta)\,k_{t+1}=(1-\delta)\,k_t+x_t, \tag{3}$$

$$z_{t+1}=\rho z_t+\epsilon_t, \tag{4}$$

where β is a time discount factor, α and $(1-\alpha)$ are Cobb–Douglas consumption shares, x_t is time t investment per head, c_t is time t consumption per head, k_t is capital per unit of effective labour, z_t is a random productivity parameter, h_t are hours supplied at time t, θ and $(1-\theta)$ are share parameters in Cobb–Douglas technology, δ is the depreciation rate, and ϵ_t a productivity process error term.

Cooley and Prescott move sequentially through this model to calibrate it. θ is capital's share in output, which from (modified) national accounts data is 0.40; labour's share $(1-\theta)$ is 0.60. Substituting the constraints into the objective function and taking first order conditions, Cooley and Prescott obtain

$$\frac{(1+\gamma)(1+\eta)}{c_t} = \frac{\beta(1+\eta)\left[\theta k_{t+1}^{\theta-1} h_{t+1}^{1-\theta} + (1-\delta)\right]}{c_{t+1}}. \tag{5}$$

So that the steady-state is a stationary distribution in which k_t, c_t, and y_t (output per effective worker) grow at the same rate (and dropping the t subscripts)[6],

$$\frac{(1+\gamma)}{\beta} + \delta - 1 = \theta \frac{y}{k}. \tag{6}$$

Along a balanced growth path, first order conditions for hours yield

$$(1-\theta)\frac{y}{c} = \frac{\alpha}{(1-\alpha)}\frac{h}{(1-h)}. \tag{7}$$

Cooley and Prescott show the laws of motion for capital stock growth imply that in a steady state

$$(1+\gamma)(1+\eta)\frac{k}{y} = (1-\delta)\frac{k}{y} + \frac{x}{y}, \tag{8}$$

or,

$$\delta = \frac{x}{k} + 1 - (1+\gamma)(1+\eta). \tag{9}$$

Equation (9) allows δ, the depreciation rate, to be determined once the growth rates γ and η are known, and data on the investment to capital ratio is selected. Once δ is determined, Equation (6) allows β (the discount factor) to be determined, given that θ (the Cobb–Douglas preference share parameter in production) has been determined from factor share data. Equation (7) allows α, the Cobb–Douglas share parameter to be determined, if h (hours worked) are known from data. Cooley and Prescott cite time use survey data referenced in Ghez and Becker (1975) as showing that households devote about one third of their discretionary (non-sleep) time to market activity, and

[6] With $0 < \rho < 1$ there will be a stationary distribution for the levels of k, c and y.

use a value of $h = 0.31$. They also calibrate to a steady-state, output to consumption ratio of 1.33, implying that $h(1 - \alpha)$ has a value of 1.78.

Cooley and Prescott complete their calibration by determining parameters of the process that generates the shocks to technology in Equation (4). They assume a value for ρ of 0.95 which, from data on measured GNP for the US, implies innovations to technology that have a standard deviation of about 0.007.

Recent literature on calibrated macro models has been concerned with related issues that these forms of calibration raise. One is how to solve recursive dynamic programming problems; den Haan and Marcet (1990) present a method of parameterized expectations; Judd (1991) uses a minimum weighted residuals technique. King, Plosser and Rebelo (1988a,b) simplify their model for this purpose by using a linear approximation to the first order conditions characterizing equilibrium, a procedure reminiscent of Johansen's (1960) linearized general equilibrium model of Norwegian growth on which micro modellers have built.

Another is how best to represent the data on business cycles which these models seek to explain. A key component of the literature has been the use of Hodrick–Prescott filters [see Hodrick and Prescott (1980)]. These are parameters, λ, which embody the relative variance of the growth component to the cyclical component in data, and are chosen so as best to trade off the extent to which the growth component tracks the actual series, against the smoothness of the trend. As $\lambda \rightarrow \infty$, the growth component approaches a linear trend. Hodrick–Prescott filters pre-adjust data, but in ways different from the micro-based policy calibration literature.

Further work discusses how best to assess the performance of calibrated models. Watson (1993) provides measures of fit for calibrated models, based on the size of the stochastic error needed to match the second moments of actual data exactly. Christiano and Eichenbaum (1992) offer a generalized method of moments interpretation of calibration exercises, which provides a metric for assessing the difference between model predictions and the data.

Real business cycle models have also been elaborated on in terms of structure and coverage. Rupert, Rogerson and Wright (1995) discuss models with a more extensive treatment of household production, a structure subsequently further explored in McGrattan, Rogerson and Wright (1997). Andolfatto (1996) discusses the qualitative implication of labour market search for fluctuations in a real business cycle model, while Coleman (1996) addresses the correlation between money and output in a model where the quantity of money is endogenously determined and, in the long run, neutral.

Although the origins of macro-type calibration lie in the real business cycle literature, these techniques have also been applied to other economic questions. One such field of application has been asset pricing. Mehra and Prescott (1985) use a calibrated model to show that under reasonable restrictions, standard competitive theory cannot explain both the low average real returns to debt and the high returns to stocks; the so-called equity premium puzzle. Other anomalies in asset prices have been uncovered using calibrated models, and these are surveyed in Kocherlakota (1996).

Calibrated models are also tools employed in the endogenous growth literature. For example, King and Rebelo (1993) assign numerical values to the parameters of a neoclassical growth model and conclude that the model's predictions are inconsistent with interest rate, asset price and factor share observations. Importantly, they find that when the initial capital stock is sufficiently low, capital accumulation makes a large contribution to growth but the marginal product of capital in the early stages of growth is unrealistically high.

Tax analysis in the real business cycle context has also been a focus of attention. Greenwood and Huffman (1991) use a calibrated model to examine the welfare effects of distortionary taxes relative to the costs of volatility in aggregate variables and conclude that the welfare costs of taxation in their model are greater than those of fluctuations over the business cycle. Cooley and Hansen (1989) incorporate cash-in-advance constraints in their real business cycle analysis. Their simulations suggest that the level of expected inflation does not affect the business cycle, although it does have an impact on long run values of aggregate variables.

Calibrated business cycle and endogenous growth models have also been used to examine fiscal policy issues. Papers surveyed in Stokey and Rebelo (1995) use endogenous growth models calibrated to US data to infer the effects of income taxes on the growth rate, and conclude variously that tax reform would have little impact on growth [Lucas (1990)], modest impact on growth [King and Rebelo (1990)] and a large impact on growth [Jones et al. (1993)]. Stokey and Rebelo undertake sensitivity analysis with respect to the numerical specification of several parameters to uncover the source of the discrepancy in these various conclusions.

4.2. *Calibrated microeconomic models*

The real business cycle macro models deal with a small number of aggregate variables and their dynamics include relatively few parameters, so that the number of observations far exceeds the number of parameters. In contrast, the approach to calibration in policy-oriented micro models is usually exact. Given the disaggregation used in these models for goods, factors and agents (producers, consumers and governments), they typically have many more parameters than can be inferred from the data. Hence, modellers exogenously specify the values for a sufficient number of parameters (usually the elasticities), so that the remaining free parameters can be calibrated to fit the data exactly. Thus, by definition the minimized distance criterion defined over the difference between data and model outcomes is zero in these models, since the data are in fact pre-adjusted to conform to the equilibrium solution concept of the model.

In calibration for policy evaluation, values for the model parameters are recovered from adjusted data in a deterministic manner. Parameters, and hence calibration of their numerical values, are thus viewed differently in real business cycle macro models and policy evaluation oriented micro models. In the former, the parameters are often the so-called "deep" parameters of technology and tastes that are viewed as unlikely to change over long periods. They are of interest on their own, and calibration is, in effect, an attempt to recover them from aggregate data. In the latter, the focus is on the comparative static or, less often, comparative dynamic effects of policy changes, on equilibria.

Micro modellers typically see their simulations largely as numerical implementations of theoretical structures. To them, the widespread use of a particular structure in the theoretical literature is an indication of its worth, so that they seek less to test or validate models and more to explore the numerical implications of a particular model, conditional on having chosen it. Policy modellers tend to be agnostic about particular models, accepting that many alternative structures relevant to an issue exist in the theoretical literature, sometimes producing different results. Thus, unlike the real business cycle modellers, the focus of micro modellers is to generate insights about the effects of policy or other changes conditional on a particular theoretical structure, rather than to test theory itself.

An illustration is provided by the customs union issue. In the 1950s, when faced with the Treaty of Rome and what today is the European Union, trade economists began to explore the implications of regional trade agreements. Following Jacob Viner's (1950) work on the customs union issue, theoretical trade economists began to debate the relative importance of the trade diversion and trade creation effects that Viner had identified as stemming from a Union. These two effects clearly operated in opposite directions, and so for any individual country contemplating joining a union, numerical calculations of equilibria were needed.

In later work, modellers developed a calibration methodology that forces model parameters to reproduce data exactly as a base case. This type of calibration, set out in Mansur and Whalley (1984) and Shoven and Whalley (1992), is widely used in micro models. It can be illustrated by considering a simple general equilibrium model with consumption and production in which two consumers are endowed with two factors of production. These factors combine to produce two goods using CES technology, and the consumers have Cobb–Douglas preferences over the two goods. Consumers' demands reflect utility maximization subject to a budget constraint.

The endogenously determined variables in this model are X_i^h, consumer h's demand for good i; Q_i, the quantity of good i produced; P_i, the price of output i; F_i^j, the use of factor j in the production of good i; and w_i the price of factor i. Here, the price of factor 2, w^2, is arbitrarily chosen as the numeraire and is set equal to 1. Consumer h's endowment of factor j, E_j^h, is exogenously given.

An equilibrium for this model is a set of goods and factor prices such that

(i) Factor markets clear:

$$\sum_i F_i^j - \sum_h E_j^h = 0 \quad (j = 1, 2), \tag{10}$$

(ii) Goods markets clear:

$$\sum_h X_i^h - Q_i = 0 \quad (i = 1, 2), \tag{11}$$

Table 1
An example of a microconsistent data set used in calibration of a simple general equilibrium model

	Production of Good 1	Production of Good 2
Production		
Value of production	20	26
Value of input use of factor 1	12	10
Value of input use of factor 2	8	16
	Consumer 1	Consumer 2
Demands		
Value of demand for good 1	9	11
Value of demand for good 2	9	17
Income	18	28
Income sources		
Value of endowment of factor 1	10	12
Value of endowment of factor 2	8	16
Income	18	28

(iii) Zero profit conditions hold:

$$P_i Q_i - \sum_j w^j F_i^j = 0 \quad (i = 1, 2). \tag{12}$$

Because consumer demands reflect utility maximizing behaviour, they will satisfy budget balance so that $\sum_i P_i X_i^h = \sum_j w^j E_j^h$ and Walras' Law holds.

A micro calibration of this model uses equilibrium data to find the values of the share parameters in the consumer utility functions and the share and scale parameters in the production functions. To be used as input data to calibration, however, data must be consistent with an equilibrium solution to the model, that is, they must satisfy the equilibrium conditions. Table 1 provides an example of such data: the value of inputs equals the value of outputs in each sector, the value of consumption equals that of production of each good, and the consumers are on their budget constraints.

Table 1 reports observations in value terms, but to undertake calibration the value observations need to be separated into price and quantity observations. This is done through the choice of units. A units convention originally adopted by Harberger (1962), and widely followed since, is that for both goods and services, quantities can be defined as those which sell for one unit of currency. This convention allows all base case prices in the economy to be set to 1. It also implies that the values of transactions in Table 1 denote quantities transacted.

The utility functions in this example are Cobb–Douglas. If U^h is the utility of consumer h, and X_i^h is consumer h's demand for good i, each consumer's utility can be written as

$$U^h = \prod_i \left(X_i^h\right)^{\beta_i^h}, \quad (h = 1, 2), \tag{13}$$

where β_i^h is the expenditure share of good i for consumer h and $\sum_i \beta_i^h = 1$.

Production functions in this model use the CES functional form so that the output of good i, Q_i, is given by

$$Q_i = \lambda_i \left(\sum_j \alpha_i^j \left(F_i^j\right)^{(\sigma_i - 1)/\sigma_i} \right)^{\sigma_i/(\sigma_i - 1)} \qquad (i = 1, 2), \tag{14}$$

where F_i^j is the input of factor j into the production of good i, λ_i is a scale parameter, σ_i is the constant elasticity of substitution, α_i^j is the CES share parameter in the production of good i, and $\sum_j \alpha_i^j = 1$.

With Cobb–Douglas demands and CES production there are twelve parameters in this model. These are for the shares for goods in each consumer's preferences β_i^h; the CES share parameters for factor j in the production of good i, α_i^j; the scale parameters in the production function for good i, λ_i; and the two elasticities of substitution in the CES production functions, σ_i.

Cobb–Douglas demands are given by

$$X_i^h = \frac{\beta_i^h I^h}{P_i} \quad (h = 1, 2; \quad i = 1, 2), \tag{15}$$

where I_h is consumer h's income and is defined as $I_h = \sum_j w^j E_j^h$. For known solutions values X_i^h, I^h, and P_i, calibrated demand parameters are given by

$$\beta_i^h = \frac{P_i X_i^h}{I_h} \quad (h = 1, 2; \quad i = 1, 2). \tag{16}$$

On the production side, the CES factor demand functions are

$$F_i^j = \frac{Q_i \left(\alpha_i^j\right)^{\sigma^i}}{\lambda_i (w^j)^{\sigma^i} \left[\sum_j \left(\alpha_i^j\right)^{\sigma^i} (w^j)^{1 - \sigma^i} \right]^{\sigma^i/(\sigma^i - 1)}} \quad (i = 1, 2; \quad j = 1, 2). \tag{17}$$

In calibrating production parameters the usual procedure is to set values for the elasticity parameters σ^i. First order conditions from profit maximization then yield calibrated factor share parameters as

$$\alpha_i^j = \frac{w^j \left(F_i^j\right)^{1/\sigma^i}}{\sum_j w^j \left(F_i^j\right)^{1/\sigma^i}} \quad (i = 1, 2; \quad j = 1, 2). \tag{18}$$

Table 2
Calibrated parameter values for Cobb–Douglas general equilibrium model using the microconsistent data from Table 1 [a]

Utility function share parameters	β_1^1	$=$	0.50
	β_2^1	$=$	0.50
	β_1^2	$\doteq$	0.39
	β_2^2	$\doteq$	0.61
Production function share parameters	α_1^1	$\doteq$	0.58
	α_1^2	$\doteq$	0.42
	α_2^1	$\doteq$	0.36
	α_2^2	$\doteq$	0.64
Production function scale parameters	λ_1	$\doteq$	1.97
	λ_2	$\doteq$	1.93

[a] $\doteq$, approximately equal to.

Substituting the α_i^j into the production function allows the calibration of the scale parameters λ_i,

$$\lambda_i = \frac{Q_i}{\left[\sum_j \alpha_i^j \left(F_i^j\right)^{(\sigma^i - 1)/\sigma^i}\right]^{\sigma^i/(\sigma^i - 1)}} \quad (i = 1, 2). \tag{19}$$

If we suppose that either econometric estimation or a literature search have yielded elasticity values $\sigma^1 = 1.2$, and $\sigma^2 = 0.8$, the calibrated parameter values using the data from Table 1 and the elasticities are given in Table 2.

If the calibrated parameter values set out in Table 2 are used in solving the model, the equilibrium solution values will be those given by the data in Table 1. This replication test is used by micro-modellers to ensure that no errors are present, either in the calibration calculations or in the model coding. One issue discussed by some modellers is the possibility of multiple equilibria, so that the replication test fails because the model solves for an equilibrium other than that of the base case data. Numerical examples of multiple equilibria have been constructed by Kehoe (1985) for simple Cobb–Douglas economies with a small number of production activities, which have caused some disquiet because of the presence of multiplicity in seemingly simple models. However, where smooth production functions of the Cobb–Douglas or CES variety are used, most modellers believe that uniqueness is the more likely outcome [see Kehoe and Whalley (1985)]. *Ad hoc* tests undertaken with applied models, seem to confirm this view. Such tests include setting the model's starting values to a slightly displaced version of the initial equilibrium solution and checking that the model calculates the initial equilibrium as a solution, and approaching equilibria at different speeds and from different starting points.

Although the example used here is simple, the same calibration approach can be followed in larger scale models. Piggott and Whalley (1985) use a general equilibrium model of the UK with 100 households, 33 productive sectors, and 29 traded goods. Including the intermediate production structure, the model uses around 20 000 parameter values. Models of these dimensions are not exceptional. An even larger model, the ORANI model of the Australian economy, identifies 115 commodities and 113 industries in its base period input–output data [Dixon, Parmenter, Sutton and Vincent (1982, p. 202)].

Micro general equilibrium models are inevitably more concerned with disaggregated representations of economies than is true of the real business cycle models. For example, the Whalley (1985a) global trade model considers four trade blocs, each of which produces 33 commodities, and has a government and several household types as consumers. This scope for accommodating detail enables modellers to address focussed policy questions which, by virtue of their disaggregated data requirements, are difficult using an econometric approach. Questions that have been examined in such a framework include the impact of the Canada–US Free Trade Agreement [Cox and Harris (1985)], the effects of the Uruguay Round of GATT negotiations [Nguyen et al. (1996)], tax incidence in Côte d'Ivoire [Chia et al. (1992)], and the potential impact of carbon taxes [Whalley and Wigle (1990)].

The data requirements of micro-based policy models can be large. In the Whalley (1985a) model, they include trade between regions, trade barriers within regions, factor endowments and consumption by household type in each region, factor inputs and the value of output for each production sector by region; all in consistent units and for a specified benchmark year. Such data are derived from several sources, including trade data, household expenditure surveys, input–output tables, government administrative records, statistics from taxation departments and national income accounts. Some information, such as some of the trade barrier data, is either missing or incomplete; and worse, key elasticity parameters need to be specified for which literature based estimates may be non-existent, or if they exist may be contradictory. A further problem is that the structure of econometric models from which elasticity parameters may have been estimated is likely to be dissimilar to that of the micro model in which they are used, since the estimation procedures are unlikely to have imposed a general equilibrium structure on the data. Hence, a precise match between collected parameter estimates and the model's input requirements for calibration seldom exists.

Reconciling and adjusting these diverse data sources so that they form a consistent data set can pose challenges, and these are set out in some detail in St. Hilaire and Whalley (1983). The levels of sectoral, household or product aggregation can differ among data sources. Definitions of terms can vary, and do not necessarily accord with the model requirements. Gaps can occur, with no estimate available for some components of the required data. Measurement errors abound; an estimate of the same variable in one data source may differ sharply from that in another. Data sources themselves also vary in their reliability because collection techniques and methods of analysis differ among researchers and institutions. Including more than one country in a data set compounds these consistency problems.

4.3. Deficiencies of micro calibration

The use of data adjustments and the absence of statistical structures in the micro models precludes the use of these models for forecasting. Despite basing conclusions on the best available data, modellers do not pretend that their model results yield anything other than indications of the relative orders of magnitude for possible policy adjustments in the economy. Often modellers seek quantitatively informed insights. Are effects large or are they small; are they opposite to received wisdom, and if so why. Paradoxically, the framework that allows a detailed specification of the economic system introduces uncertainty into the model conclusions by virtue of its requirement for highly disaggregated and, inevitably, approximate data.

The dimensionality of these models is another issue. In some circles, due to their detail, micro models have developed reputations as black boxes into which a policy change is fed as an input and from which a set of results emerges with little explanation. In these models, the key interactions that drive the model results can easily become obscured. The data requirements of a detailed model can also be a constraint in the implementation of calibration. Model detail often centres on the sectors and agents most likely to be affected by the policy question at issue, while the remainder of the economy is modelled at a relatively more aggregated level.

The calibration of such models also implies no model testing. Many different models with different structures could, in principle, be calibrated to the same data set. Modelling efforts are seen as solely theory with numbers, the aim of which is to provide model conditional insights, such as which effects are large and which small, which positive and which negative; either for policy input or for the better understanding of economic processes. Typically, model structures are sufficiently complex that estimating their parameters in a form which imposes the equilibrium solution concept of the model is infeasible. Econometrically estimated models generate conditional inferences, since they maintain assumptions about stochastic distributions, but their statistical basis allows greater rigour to be attached to insights. In contrast, results from micro simulation models are not based on any formal statistical criteria since they do not aim to provide forecasts. Model results are suggestive, and their interpretation subjective.

4.4. Some recent examples of calibrations

Far from being a passing trend, calibration continues to spread as a numerical technique used in economics. Some recent examples of calibrated models, together with summaries of their structures, calibration procedures, data sources, key parameters and main conclusions are given in Table 3. As this Table indicates, the distinction between macro-oriented and micro-based models persists in current modelling practice. Macro models are based on a dynamic growth structure and employ calibration techniques which have their foundations in the Kydland and Prescott calibration, whereas the static, general equilibrium models continue largely to calibrate parameters following the approach from Shoven and Whalley.

The four recent dynamic models in Table 3 exhibit greater variation in their approaches to calibration than do the three static models; Greenwood, Hercowitz and Krusell (1997) use a dynamic growth model to examine the quantitative effects of technological change in equipment, which is investment specific, on US growth, and conclude that it accounts for a larger proportion of growth than does neutral productivity change. Their simulations are undertaken using a representative consumer model with Cobb–Douglas production. Investment-specific technological change is introduced in the evolution of the stock of equipment, one of the factors in the production function; the contribution of investment to the changing stock of equipment is scaled by a technology parameter. The technology parameter is modelled as a random process with a known average growth rate which is calculated from an exogenous data series. Depreciation rates and marginal tax rates of labour and capital income are likewise assigned from non-model sources. The remaining model parameters, including preference parameters and factor shares in production, are calibrated so that the model reproduces features of long-run US data.

Cooley, Hansen and Prescott (1995) examine the effects of idle resources on business cycle fluctuations in the US. They also use a dynamic growth model. Production in their model is undertaken in a continuum of plants that vary in both location and stock of physical capital, where each plant is indexed by a technology parameter added to a random shock. Below some threshold value of the technology parameter, a plant is no longer profitable and will shut down – a mechanism which allows for the possibility of idle resources in the economy. The value of the technology parameter is set in one case to mimic the standard business cycle model, and in other cases to match literature based data on average capacity utilization. The remaining parameters are calibrated so that the non-stochastic steady-state behaviour of the model matches features of US data.

Huggett (1996) uses an overlapping generations model where consumers differ in their stage in the life-cycle, to examine how models with uncertainty in earnings and lifetimes match US data on the age-wealth distribution. In contrast to the two previous papers, the focus of the model structure is on the consumer rather than the producer specification. Calibration of the model relies heavily on values taken from other published sources and includes actuarial estimates of survival probabilities for different age groups, published population growth and labour force participation rates, previous estimates of relative risk aversion and depreciation. The age–earnings profile is calibrated from data for the US.

The calibration in the Huggett model uses literature-based parameter values almost exclusively, whereas the remaining modellers choose some parameters so that the model outcome matches data and draw others from the literature. This use of the term

Table 3
Calibration procedures used in some recent modelling exercises

Reference	Model focus	Model structure	Calibration procedure	Primary data sources	Key parameters	Main conclusions
Greenwood, Hercowitz and Krusell (1997)	Role of investment-specific technological change on growth in the US.	Dynamic growth model with vintage capital in production.	Growth rate of new equipment productivity, depreciation rates, and labour tax rates matched to literature values. Remaining parameters calibrated to match features of US data.	National Income and Products Accounts for the US. Literature estimates for specific parameter values.	Productivity of a new unit of equipment changes over time and is given by direct observation.	Investment-specific technological change accounts for about 60% of growth in the US over the post-war period.
Cooley, Hansen and Prescott (1995)	Effects of idle resources on business cycle fluctuations in the US.	Dynamic growth model with firms which have plants that are defined as a location and physical capital. Technology shocks determine the fraction of idle capital in each period.	Parameters calibrated to be consistent with capital-output ratio, investment share in output, the share of income paid to capital and the average share of time in market labour for the US.	National Accounts Data for the US. Literature estimates for specific parameter values.	Average rate of capacity utilization taken from literature. Consumption parameters and depreciation calibrated to long run growth values.	Inclusion of variable capacity utilization and idle resources does not affect cyclical properties of the model significantly, but does give variation in factor shares.

continued on next page

Table 3, *continued*

Reference	Model focus	Model structure	Calibration procedure	Primary data sources	Key parameters	Main conclusions
Huggett (1996)	How life-cycle economies with uncertainty in earnings and lifetime match aggregate and transfer wealth in the US.	Overlapping generations model with shocks to labour productivity.	Parameters set to match values from the literature.	Various literature estimates. Handbook of Labor Statistics, Social Security Bulletins.	Preference, technology, demographic and tax parameters.	Including earnings and lifetime uncertainty in lifecycle models replicates aggregate and transfer wealth distributions for the US.
Parente and Prescott (1994)	Role of barriers to technology adoption in cross-country income disparity.	Dynamic growth model where firms invest to improve technology, but results are hampered by barriers. Behaviour of Japanese and US modelled economies differs only in barriers to technology adoption.	Parameters calibrated to US balanced growth path are contingent on a technology capital share parameter derived from postwar Japanese development. Several numerically feasible values for this parameter exist, but the value is chosen on plausibility criteria.	US national income and product accounts. Literature values for specific parameter values.	Parameter for barrier to technology adoption normalized to 1 for US. Simultaneous growth rates and business physical capital stock from literature. Depreciation, tax rates, consumption parameters and rental rates are calibrated from data.	Development miracles and the disparity in cross-country incomes can be reasonably explained by changes in barriers to technology adoption.

continued on next page

Table 3, *continued*

Reference	Model focus	Model structure	Calibration procedure	Primary data sources	Key parameters	Main conclusions
Various GTAP[1] models in Hertel (1997)	Various including the benefits of abolishing the MFA, effects of climate change on agriculture, issues on multilateral vs preferential free trade in the Pacific Rim.	All are adaptations of the static, 24 regions, 37 sectors GTAP global trade model.	Aggregation of general GTAP database to meet specific model focus. Calibration of shares is to 1992 levels data and elasticities to changes data.	GTAP multiregional database derived from individual country input–output tables, bilateral trade data, and Uruguay Round GATT data.	Elasticities of substitution and income elasticities derived from literature. Constant difference elasticity function calibrated to approximate own price elasticities.	Various conclusions including: MFA reform leads to large global welfare gains; including the link between CO_2 and crop growth in models mitigates estimates of the damaging impact of climate change on welfare; reciprocity of non-APEC members for nonpreferential trade liberalization affects whether benefits are conferred on APEC or non-member countries.
Piggott and Whalley (1998)	Role of self supply and underground economy in analysis of VAT options in Canada.	Static, 2 consumers, 3 goods, single region model. Includes self supply and underground production.	Calibration to 1993 levels data and changes data following 1989 Canadian tax changes.	National accounts data, Time Use Survey, tax data, combined into single consistent data set.	Supply elasticities inferred from changes data. Literature based demand elasticities. Calibrated share parameters.	The presence of self-suppliable goods can reverse the result that broadening the VAT base is welfare improving.

continued on next page

Table 3, *continued*

Reference	Model focus	Model structure	Calibration procedure	Primary data sources	Key parameters	Main conclusions
Lee and Roland-Holst (1997)	Environmental implications of tax and trade policies in Indonesia.	Static, 3 region, 19 goods model. Includes 10 pollutants from production and emissions taxes.	Calibrated to 1985 levels data for Indonesia and Japan.	Industrial Pollution Projection System database used. Input–output matrix, employment and capital stock data combined into single consistent data set.	Taxes and tariff share parameters calibrated from data set. Effluent coefficients calibrated from pollution database.	A uniform effluent tax is the most cost efficient option for abating emissions and its use together with trade liberalization can counteract the negative environmental effects of output growth.

[1] GTAP is the acronym for the Global Trade Analysis Project, a project designed to facilitate the general equilibrium analysis of global trade issues.

calibration to include parameters drawn from literature contrasts with the use of the term in micro-based models where a model's "calibrated" parameters are restricted to those derived from the benchmark data set and the literature-based parameters such as elasticities are considered parameters with assigned values. Including assigned parameter values that have been derived outside the model structure under the calibration rubric is a potential source of confusion in the communication between micro and macro calibrators.

In contrast to Huggett (1996), Parente and Prescott (1994) employ a more model-dependent approach to calibrating parameters. They use a dynamic general equilibrium model to show how barriers to the adoption of technology in different countries lead to differences in per capita income. Data on the US and Japanese economies form the basis of their calibration. In their model structure, firms in each country are differentiated by their technology levels. To move from one technology level to another, firms must invest. The technological advancement from a particular level of investment, however, depends crucially on two parameters – one relates to country-specific technology adoption barriers (π), and the other, θ_z, which is not country-specific, relates to the firm's current level of technology relative to world knowledge. In the case of the US, the value of π is set to 1. Calibration of model parameters is undertaken with respect to long run balanced growth data for the US. Such calibration requires, however, the specification of a value for θ_z. This value is found by calculating (π, θ_z) pairs for the Japanese economy, finding the implied income level relative to the US for each pair, and choosing the pair which is most consistent with data.

Table 3 also summarizes how calibration has been used in three recent static applied general equilibrium modelling exercises relating to trade, tax and environment issues. The static models exhibit greater variety in dimensions than their dynamic counterparts, but offer a more uniform approach to calibration. The large scale, highly disaggregated Hertel (1997) models, which are the product of long-term modelling efforts, are designed to be sufficiently realistic representations of the global economy that they can be used for policy analysis. They contrast with the small, focussed model of the Canadian economy in Piggott and Whalley (1998) which is designed to investigate a specific proposition about tax reform.

The models included in the Hertel (1997) volume focus on a range of issues related to international trade, including the effects of abolishing the Multi Fibre Arrangement [Yang, Martin and Yanagishima (1997)], the effects on relative wages in industrial countries of developing country expansion [McDougall and Tyers (1997)], the effects of global climate change on agriculture [Tsigas, Frisvold and Kuhn (1997)], and a comparison of multilateral and preferential free trade in the Pacific Rim [Young and Huff (1997)]. All of these models use the data base developed under the Global Trade Analysis Project (GTAP) as a basis for calibration, although each of the issues addressed requires the modellers to aggregate and adapt the basic GTAP model and data framework to match the problem under consideration.

The disaggregation of the GTAP data base is the feature that allows it to form the basis of such a variety of trade oriented policy analyses. Perhaps more than any of the other models listed in Table 3, the GTAP models illustrate how calibration has given policy modellers the flexibility to address issues which would be infeasible in an econometric framework. The GTAP data base is large, identifying production, consumption, trade flows and tariffs among 24 regions each with 37 production sectors. The derivation of the single 1992 observation, which comprises the data base, was itself a large undertaking. Obtaining sufficient time series to estimate the model parameters would be close to impossible.

In contrast, Piggott and Whalley (1998) calibrate a small, 2 consumer, 3 goods model in which labour is the only factor of production. They use the model, calibrated to Canadian data, to show how the presence of self-supplied goods and the underground economy can reverse the conventional wisdom that broadening the base of a VAT is welfare improving. The direction of welfare change in their model hinges crucially on the supply elasticity for non-market goods. Instead of using literature-based estimates or using a "best guess" approach as is the traditional approach for finding elasticity parameters in micro-based models, they use information on the actual response of the Canadian economy to the imposition of the VAT to calibrate its value.

The disaggregation of the GTAP models and focussed approach of the Piggott and Whalley model represent two ends of the spectrum of micro models. The dimensions and structure of the final model in Table 3 are perhaps more representative of the majority of micro-based, static models. Lee and Roland-Holst (1997) use a standard general equilibrium framework with perfect competition, constant returns to scale production technologies, fixed stocks of capital and labour, and endogenous trade flows to examine the environmental implications of tax and trade policies in Indonesia. They deviate from the standard model by including detailed specification of sectoral emissions, in which emissions are linear in sectoral output. Data from a pollution database is used to calibrate the effluent coefficients.

5. Best practice in calibration

Given the widespread use of calibrated models in modern economics, the question of what comprises best practice naturally arises. In econometric work, debate centres on estimation procedures, the properties of estimators, the appropriateness of tests and the development of test statistics, as well as other issues of implementation. With calibrated models, authors have largely been content to describe their model parameterization procedures by appealing to the term calibration, giving few or sometimes no details about their procedures. To our knowledge, no discussion exists in the literature as to whether one set of calibration procedures is to be preferred to another.

5.1. The choice of model for calibration

Perhaps the major Achilles' heel in the use of calibrated models for empirical investigation is the choice of model, both because models are not tested against one another, and because the precise model form can have a major influence on results. Reference to widely used theoretical structures is usually an insufficient basis on which to choose models, especially since much of the theoretical discussion is oriented towards showing how changes in model structure can often change the qualitative model predictions. Model selection based on theoretical literature may sound appealing, but the literature does not offer guidance on the precise specification of the model to be used, nor does it provide the criteria under which such a choice should be made.

An example of how the conclusions of calibrated models can change substantially with model structure may help to illustrate the point. In 1962, Harberger performed some of the earliest general equilibrium simulations, implicitly calibrating a two sector model of the US economy and evaluating counterfactuals to show that a tax on one factor in one sector (the tax on capital in the corporate sector) was borne fully by that factor even if it was mobile between the two sectors. In fifteen years of subsequent literature, the addition of more sectoral disaggregation, partially mobile factors and other features failed to change the basic result; capital still bore the burden of the corporate tax.

In the late 1970s, however, simulations showed that if the US economy were modelled as facing a perfectly elastic supply function for capital, instead of the fixed endowment, inelastic supply function, assumption of Harberger, this result would reverse. Capital could not bear the burden of the tax in such a situation, and it must be shifted elsewhere. Two model structures could yield this feature – one with perfect international capital mobility, or one with an intertemporal structure with savings (consumption smoothing) where the savings elasticity and hence, the supply elasticity for capital within a period, is high. Modifying the original Harberger structure in either of these two directions changes the essential result.

Ambiguities in the theoretical literature revealed that even qualitative results depend upon assumptions and these are not avoided by merely using a calibrated model. Calibrating a model gives no guide as to how to choose it. The model serves as a maintained hypothesis as in econometric analysis: it enables calibration and subsequent policy evaluation, but these are conditioned on the model chosen, just as the choice of tests and inferences for econometric models are conditioned on the chosen maintained hypotheses. Unfortunately, for calibration procedures, no analogue exists to the econometric testing of alternative models; nor to the more general model choice techniques such as those employed by Pesaran (1974). Model choice for calibration remains a subjective judgement call.

Does a best practice exist for the model choice element of calibration? As long as calibrators reject the notion of model testing, and see model selection as based on a reading of theoretical literature, objective criteria in this area may be unattainable. Modellers can, however, more forcefully state that all their results are conditional upon the choice of model. They can identify which features of their model results are sensitive to which assumptions from theoretical literature, and then modify these assumptions to assess numerical sensitivity. Developing this direction in calibration allows modellers to explore the structural sensitivity of model results[7].

5.2. *The calibration of what to what?*

Even if the model to be used has been selected with sound judgement, the issue remains in calibration of what is calibrated to what. Typically, a complete specification of all model parameters is not determined through calibration; a subset of parameters is determined in this way, while a further subset is not. Calibration also sometimes only involves a partial determination of model parameter values. Some modellers appeal to the word calibration to legitimize what appears to be an *ad hoc* setting of model parameter values, with some based on literature values and some on intuition. Which parameters are calibrated is crucial for the usefulness of the modelling exercise.

As an example, in the micro-based modelling areas of public finance and trade, the conventional approach is to calibrate models to a single, constructed, equilibrium observation, as discussed above. We term this approach a "levels" calibration; once calibrated the model reproduces the equilibrium observation as a full solution to the model. However, if the model is to be used for comparative static purposes, the important component in the model specification is how the behaviour in the model changes as policy parameters or other exogenous parameters change. To the extent the calibrated level parameters, which are typically the share parameters in preferences and technology, are deemed "deep" parameters à la Lucas, this approach may seem sensible since these parameters are unaffected by policy changes. The elasticities, however, rather than these deep parameters, typically determine the comparative static behaviour of the model. The Lucas critique of econometric modelling applies to calibration here, since with the setting of elasticities or other assigned parameters, parameters drawn from a diverse literature are likely to be dependent on the policy framework applicable to the data from which they were estimated.

However, the choice of parameters to be calibrated should also depend upon the question to be asked with the model. If the issue is the welfare effects in money metric terms from a policy change, both levels parameters and changes parameters will influence the size of welfare effects. If the issue is welfare effects as a proportion of GDP, elasticity parameters are the more crucial parameters. Hence, the value of a particular calibration is a function of the particular question which the model-driven research seeks to answer.

The use of econometric estimates of elasticities from literature in calibrated models faces several problems. Large gaps exist in the literature; classifications and commodity and industry definitions between literature and models can differ widely; and where present, estimates may be contradictory. In some calibration exercises, modellers

[7] Structural sensitivity analysis is distinct from the parametric sensitivity analysis that we discuss later.

use both levels and changes elements in their calibration; in effect substituting the calibration of elasticity related parameters for the use of literature based econometric estimates.

In recent work on the effect of a sales tax reform in Canada in 1990, Piggott and Whalley (1998) note that the use of their model is not for counterfactual analysis, since data are available both before and after the reform. They use a model which, unlike others, incorporates self supply and underground economy features, and in which the effect of tax reform on welfare is ambiguous. Through a double calibration to data from before and after the reform, they are able to infer preference and other model parameters, and on this basis determine whether the tax change was welfare worsening or welfare improving. In the process, they calibrate both levels and changes parameters to more than one data observation.

5.3. The choice of functional form

The issue of best practice in calibration also arises with the choice of functional forms in a model. Modellers typically follow the family of so-called convenient functional forms for which the solutions to optimization problems can be obtained analytically. Demand functions corresponding to the maximization for Cobb–Douglas, Constant Elasticity of Substitution (CES), and Stone–Geary, or Linear Expenditure System (LES), functions are commonly used.

Cobb–Douglas functions, which are the simplest, have the unfortunate properties that uncompensated own-price elasticities of demand are unity, that uncompensated cross-price elasticities are zero, and that all income elasticities of demands are unity. In contrast, CES functions relax the unitary uncompensated own-price and zero cross-price elasticities, but do so only by adding an additional parameter to the functional form relative to the Cobb–Douglas case. Typically, modellers have literature-based or other elasticity estimates to which they wish to calibrate their models. Where more than one of these elasticities exists, extra parameters are added through nested CES functions, with additional elasticity parameters entering at the various levels of nesting. A key but relatively unknown result about elasticities is that "if the demand functions are such as to satisfy the budget constraint with strict inequality ... constant price elasticities can only assume the values -1 for all own-priced elasticities and 0 for all cross-price elasticities" [Koopmans and Uzawa (1990, pp. 3–4)]. Thus, in this case assuming constant elasticities is equivalent to assuming a Cobb–Douglas utility function.

However, both CES and Cobb–Douglas preferences are homothetic and yield demand functions that have unitary income elasticities. If income elasticities are thought to be significantly different from unity some other functional form is needed, and a Stone–Geary/Linear Expenditure System with a displaced origin for utility measurement is commonly used. The minimum requirements in such a system, which can be combined with either Cobb–Douglas or CES, are typically calibrated so as to reproduce literature estimates of income elasticities of demand in the neighbourhood of the base case equilibrium.

Some modellers have moved beyond this class of convenient functional forms to use variants of flexible functional forms, typically trans-log. The basis for rejecting the convenient forms revolves around the empirical results of econometric studies which reject the separability implicit in Cobb–Douglas and CES functions. The major drawback to using more flexible functional forms is that they are not always globally convex. Because the policy changes analyzed in many models can lead to counterfactual equilibria that are far from the initial equilibrium, the use of globally convex functions is often necessary to compute a model solution.

How do these considerations translate into best practice in the choice of functional forms? Obviously a trade-off exists between simplicity and realism; but other considerations such as computational feasibility also enter the choice. As with the choice of model structure, the choice of functional form should be influenced by the issue to be investigated.

For example, a trade model which explores the claimed long-term decline in the terms of trade of commodity-exporting developing countries, and builds on the argument from Prebisch (1962) and Singer (1950) that developing country exports are necessities and their imports (capital goods) are luxuries, will need to incorporate income elasticities of demands different from one. This is because the model needs the feature that growth in both the developed and the developing countries will adversely affect the developing country's terms of trade. This feature follows directly if developed countries have income elasticities of import demands less than one while developed countries have values that are greater than one. Using models with either Cobb–Douglas or CES preferences will not meet this requirement, and a different functional form is needed. On the other hand, if the income effects from the change considered in the model are thought to be small compared to the relative price effects, a model with homothetic preferences may suffice.

Another illustration also serves to make the same point. In analyzing tax preferences towards housing, the knee-jerk reaction would be that a general equilibrium model must be superior to a partial equilibrium model. But if the literature clearly points to an own-price elasticity of say, 0.5, a partial equilibrium analysis built around this parameter value would almost certainly attract more confidence than a general equilibrium model using Cobb–Douglas preferences.

Best practice thus involves selecting functional forms by considering the uses of the model, and deciding on which simplifications are acceptable for the purposes of the analysis, the complexity of the analysis, and the model's solvability. Violations of best practice arise with the misapplication of models, and when the model's performance is clearly contrary to the empirical literature.

5.4. The use of elasticity parameters

The quantity and quality of literature-based elasticity parameter estimates for use in calibrated models is another Achilles' heel of calibration. Some years ago, one of us [Whalley (1985b, p. 27)] noted that

> It is quite extraordinary not only how little we know about numerical values of elasticities, given the significance that we attach to these in introductory courses in Economics, but how the little we think we know changes as quickly as it does. In the savings area, for instance, 10 years ago, elasticities were thought to be small, five years ago they were thought to be large, and now once again they are thought to be smaller. For many years labour supply elasticities were thought to be small, and now they are in the process of being revised upwards. In the international trade area researchers commonly use import price elasticities in the neighbourhood of unity, even for small economies, even though elasticity estimates as high as nine appear in the literature. In many areas elasticity estimates differ in both size and sign, while for a number of the issues in which applied modellers are interested in, no relevant elasticity estimates exist. The choice of elasticity values in applied models is therefore frequently based on contradictory, or little or no empirical evidence. This obviously undermines confidence in model results.

Unfortunately, little has changed in the intervening years. Faced with a relative absence of elasticity estimates, somewhat arbitrarily assigned low and high values (sometimes with a mid-range) are often used. A low value of 0.5 and a high value of 2.0 seem to be popular choices. Elasticities estimated for different classifications are routinely adopted for model use, so that for example, an estimate for the demand elasticity for food might be used to provide the demand elasticity for cheese, even though inter-food substitution is a key feature in the model. Where estimates are deemed implausible, such as trade elasticities, they are often either ignored, or arbitrarily scaled, sometimes by as much as 50%.

Faced with all this arbitrariness, it is hardly surprising that micro modellers, at least, stress that the value of their modelling results lies in providing insights, rather than point estimates or forecasts. They use the model results to answer broad questions. What are the relative magnitudes of effects? Do the results confirm or conflict with prior thinking and if so why? If no previous studies of an effect exist, what might be an initial estimate? The theme, somewhat paradoxically, is that qualitative insights are derived using a quantitative approach.

Users of real business cycle models face fewer of these problems because they are less interested in comparative statics and can restrict themselves to simple functional forms, such as Cobb–Douglas, for which these elasticity issues do not arise. This approach, of course, raises the inevitable question of why such restrictions are employed, when at a more micro level there seems no reason that they should hold.

Faced with this situation, why do policy modellers continue with their work? The response is usually that to contribute to debate on the social issues of the day a modeller must make the best of the available information, rather than refraining from any analysis until every parameter is definitively tied down. Model use is not testing in the Friedmanian/Popperian positivist tradition, but instead is a way of harnessing available information to contribute to policy making by raising the level of debate as in the Lindholm tradition of policy sciences.

Modellers such as Mansur and Whalley (1984) have even suggested that model use has generated a demand for parameter values, particularly elasticities. They advocate a reorientation of empirical work in economics away from hypothesis testing towards parameter generation, an activity which currently yields little professional reward. Mansur and Whalley even go so far as to suggest establishing an elasticity bank, to archive and grade estimates, and make them more widely available. But as noted earlier, given that these estimates are often generated from models with structures that are different from the ones the user of the elasticities imposes, this suggestion may not be practical.

Elasticity parameters in calibrated models, and especially the micro, policy-oriented models, are key parameters for model results since they are crucial in determining comparative static behaviour in models. The elasticities at issue are typically own- and cross-price elasticities and income elasticities. Importantly, in models based on a single levels calibration, these parameter settings are not endogenously determined by calibration. Instead they are typically set, either with some form of literature justification or by an appeal to intuitive plausibility.

The current situation with elasticity estimates for use in calibrated models is poor. No estimates exist for large areas of elasticities such as, for example, production functions in service sectors. There are other areas where multiple but at times contradictory estimates exist within wide ranges. Furthermore, classifications in models do not necessarily match those from which the literature-based values are derived. Modellers refer to the "idiot's law of elasticities" where all elasticities are one until someone shows them to be otherwise, or "coffee table elasticities" where informal discussions and opinions around the coffee table determine whether a value of, say, 0.5 or 2.0 is chosen.

The number and range of surveys of elasticity estimates remain surprisingly small. Table 4 reports estimates of demand elasticities by product used by Piggott and Whalley (1985) in discussing their model parameterization, which still remain widely referred to by other modellers. These draw on literature estimates which they classify by estimation method and by product. Other surveys exist in the trade area for import and export demand elasticities [Stern, Francis and Schumacher (1976)], and on the production side [Caddy (1976)]. Few such surveys exist in the recent literature, indicating, in part, the relatively small professional pay-off involved with parameter generation.

Browning et al. (1999) provide an extensive overview of some of the problems associated with using elasticity parameter values from microeconometric studies in dynamic macro models, and several of these are worth highlighting. One difficulty with using parameters from microeconometric studies in macro models is the mismatch between definitions of the parameters in the two types of models. Browning et al. use a simple dynamic model to illustrate how a broadly inclusive term such as "labour supply elasticity", for example, can encompass several distinct concepts.

Table 4
Central tendency values for own-price elasticities of household demand functions used by Piggott and Whalley (1985) in their U.K. tax model[1,2]

Industry	LES estimates	Log-linear demand estimates	Other	Total
Agriculture and fishing	0.334 (17, .03)	0.420 (25, .05)	0.562 (44, .08)	0.468 (86, .07)
Coal mining	–	0.321 (1, 0)	1.265 (2, .01)	0.950 (3, .76)
Other mining and quarrying	0.425 (1, 0)	0.905 (3, .06)	0.257 (2, .01)	0.609 (6, .13)
Food	0.353 (15, .03)	0.580 (30, .19)	0.476 (27, .08)	0.494 (72, .13)
Drink	0.617 (5, .07)	0.780 (12, .25)	0.464 (15, .06)	0.607 (32, .16)
Tobacco	–	0.611 (8, .15)	0.431 (11, .04)	0.507 (19, .10)
Mineral oils	0.425 (1, 0)	0.905 (3, .07)	0.257 (2, .01)	0.609 (6, .13)
Other coal and petroleum products	1.283 (2, .01)	1.404 (3, .80)	1.978 (3, 1.41)	1.589 (8, .90)
Chemicals	0.685 (1, 0)	0.890 (1, 0)	0.680 (3, .07)	0.724 (5, .05)
Metals	–	1.522 (19, .42)	0.989 (18, .40)	1.083 (51, .48)
Mech. engineering	–	1.296 (16, .61)	1.068 (15, .43)	1.005 (45, .48)
Instr. engineering	0.606 (14, .15)	1.099 (17, .57)	1.240 (11, .54)	0.972 (42, .49)
Elec. engineering	–	1.388 (19, .377)	1.049 (17, .41)	1.060 (50, .44)
Vehicles	0.606 (14, .15)	1.137 (19, .55)	1.099 (18, .40)	0.985 (51, .44)
Clothing	0.277 (16, .03)	0.491 (26, .16)	0.564 (19, .15)	0.458 (61, .18)
Timber, furniture, etc.	0.570 (14, .09)	1.258 (19, .23)	0.974 (20, .39)	0.969 (53, .33)
Paper, printing, publishing	0.191 (1, 0)	0.343 (5, .02)	0.416 (5, .02)	0.362 (11, .02)
Other manufacturing	0.578 (14, .02)	0.527 (7, .11)	0.626 (17, .12)	0.592 (38, .09)
Gas, electricity, water	1.203 (1, 0)	0.921 (9, .02)	0.369 (10, .01)	0.659 (20, .10)
Transport	0.761 (4, .23)	1.027 (14, .26)	0.994 (10, .16)	0.977 (28, .23)
Banking and insurance	–	0.559 (3, .02)	0.894 (1, 0)	0.642 (4, .04)
Housing services (private)	0.461 (15, .11)	0.550 (29, .45)	0.434 (9, .09)	0.505 (53, .29)
Professional services, other services	0.488 (7, .08)	1.090 (16, .39)	0.946 (16, .39)	0.961 (50, .48)

[1] All uncompensated own-price elasticity estimates; figures in parentheses refer to the number of studies included and the variance of the estimate.
[2] Source: Piggott and Whalley (1985).

They specify a single agent model with a labour–leisure choice so that at time t, the agent is assumed to choose non-durable consumption c_t and hours of work, $h_t = T - l_t$, where l_t is the choice of leisure at time t. While preferences are assumed to be intertemporally separable, utility at time t, $U(c_t, h_t)$ is not.

The labour supply function at time t, $h(\cdot)$, can be expressed as function of the price of consumption, p_t, the wage rate, w_t, and the marginal utility of income, λ_t:

$$h = h(p_t, w_t, \lambda_t). \tag{20}$$

Browning et al. proceed to derive three distinct labour supply elasticities. The first, the Frisch elasticity, θ, *conditions* on the marginal utility of income – that is, it holds λ constant. Dropping the t subscripts, it is given by

$$\theta = h_w(p, w, \lambda)(w/h), \tag{21}$$

where h_w denotes the derivative of $h(\cdot)$, the labour supply function in Equation (20) with respect to the wage rate. This elasticity incorporates the intertemporal response to wage changes.

The remaining two elasticities capture within period responses only. Browning et al. define the total net expenditure within a period as $e = pc - wh$, and then show that the uncompensated labour supply function, denoted here by $h^*(\cdot)$ can also be expressed as

$$h = h^*(p, w, e), \tag{22}$$

implying a second labour supply elasticity, θ^*, which is conditioned on e:

$$\theta^* = h_w^*(p, w, e)(w/h). \tag{23}$$

Finally, they show that the labour supply function, $h^{**}(\cdot)$, can be derived as a function of current consumption:

$$h = h^{**}(p, w, c), \tag{24}$$

and, hence, the third labour supply elasticity, θ^{**}, is conditioned on consumption and is given by

$$\theta^{**} = h_w^{**}(p, w, c)(w/h). \tag{25}$$

In general, the values of these three labour supply elasticities will differ. Because θ^* and θ^{**} are within-period responses, their values provide lower bounds for the intertemporal elasticity, θ. The handful of estimates of the intertemporal elasticity of labour supply and the static elasticity estimates given in Browning et al. support this relationship.

To be used in macro models, the econometric estimates of labour supply elasticities should be conditioned on either consumption, expenditure or the marginal utility of income, and the calibrator must ensure that the conditioning variables in the elasticity estimation coincide with the use of the parameter in the model. Browning et al. indicate, however, that many of the static microeconometric estimates of labour supply condition on none of the three relevant variables and are thus not suitable for use in calibrated macro models.

Labour supply elasticities are, of course, not the only parameters for which this issue of consistency of definitions arises. Within the context of the Browning et al. model, the consumption elasticities are also subject to conditioning variables, but the point is more general – good calibration practice demands that the conditions under which parameter values are estimated match the conditions under which they are used in the model.

Browning et al. (1999, p. 127) also argue that calibrators should "... build the dynamic economic models so that the formal incorporation of microeconomic evidence is more than an afterthought." Microeconomic evidence suggests, for example, that considerable demographically-based heterogeneity exists in parameter values relating to agents' preferences, constraints, labour supply and human capital accumulation, which is not included in the calibrated models. Empirical evidence summarised in their paper strongly suggests heterogeneity in parameter values for the discount rate and the elasticity of intertemporal substitution in consumption. For example, the discount rate for consumption has been shown to vary with household size [Zeldes (1989)], and income level [Lawrance (1991)].

Browning et al. also highlight the choice of functional forms as area in which macro model calibrators can incorporate micro evidence in the specification of their models. Many macro models rely on Cobb–Douglas functional forms which, as Browning et al. note, may be consistent with the constant capital share in the US economy, but not necessarily consistent with other observed phenomena. In general, the reliance on additively separable functional forms in macro models is not empirically substantiated. Browning et al. point to several econometric studies, including Attanasio and Weber (1993), and Attanasio and Browning (1995) to illustrate this point.

Best practice with respect to the use of microeconometric parameters in calibrated macro models thus also requires calibrators to verify the consistency between the definitions of the estimated parameters and the model specification. Despite the large number of microeconomic studies, this consistency requirement leads Browning et al. (1999, p. 127) to conclude that "... the shelf of directly usable numbers [is] virtually empty."

Given the dearth of appropriate parameter estimates, how then should the dynamic general equilibrium macro modellers proceed? The prescription we suggest mirrors that for micro calibrators. Where possible, calibrators should be economical in the number of parameters used in their model specification. Where uncertainty surrounding the parameter values exists, sensitivity analysis is appropriate. Canova (1995) presents a methodology for undertaking such sensitivity analysis with respect to the distribution of parameter values in calibrated macro models.

6. New directions in calibration

Although calibration is well established in model-based quantitative work, it is by no means a static set of procedures. In current work, new directions are evident, both in the problems to which calibrated models are being applied, and in the scope of applications which fall under the calibration rubric. The recent use of calibration techniques in *ex post* analysis, which decomposes a gross change into its constituent causes, is one example. Here, the use of two data sets, one for an initial year and one for a terminal year, allows modellers to undertake double calibration subject to constraints on which parameters may or may not be allowed to change over time. In such applications calibration is moving ever closer to estimation. Another set of developments expands the narrow definition of calibration from deriving parameters to replicate data, to analyzing the sensitivity of model results to both the parameter values and to the data adjustments undertaken in their derivation.

6.1. Expost decomposition and double or multiple calibration

The traditional use of micro-based calibrated models is to assess, *ex ante,* the potential impact of a policy change. The model's parameters are calibrated so that they replicate a base year equilibrium. Insights about the effects of a prospective policy change are then derived by introducing the policy change into the model via a parameter change, solving the model and comparing the resulting solution to the base year solution. Recently, however, the use of these types of models has spread to decompositional analysis where the modeller wishes to decompose the individual impacts from a series of simultaneous shocks to an economic system. Abrego and Whalley (1998) represents an application of these techniques to the trade and wages debate; the controversy over whether the increased dispersion of skilled versus unskilled wage rates in the US is due more to the influence of trade shocks or to technological change.

This form of analysis differs from *ex ante* policy analysis in so far as the shocks affecting an economy and the outcome of those shocks are observable in principle, even though data problems may make their joint effects imprecisely known in practice. Double calibration exploits knowledge of the outcome under the joint shock to analyze the effects of each component separately. To undertake such analysis, the modeller uses both the *ex ante* and the *ex post* equilibria and attempts to identify the effects of each component over the interval between the two. If the double calibration is exact, it identifies and finds values for parameters that are exogenous in single period calibration, such as technological parameters, or, as in the case of Piggott and Whalley (1998), elasticities which are consistent with the observed changes over time.

If the model's elasticities are derived exogenously from the literature, as in single period calibration, two period calibration can only be exact if all model parameters in technology and preferences are free to vary across time. Typically, modellers may wish to impose a constraint that technology and/or demand parameters remain unchanged over time. In this case the calibration will be inexact, and is undertaken by applying a criterion, such as minimizing the sum of squared deviations between predicted and actual variables in the two periods, to the data. Such a least squares minimization criterion, employed in Abrego and Whalley (1998), moves calibration closer to estimation.

Analyzing the effects of a specific shock, such as technological change, can then be undertaken either by allowing all of the other shocks to the economy to occur and solving the model in the absence of the shock of interest, or by only introducing the single shock to the calibrated *ex ante*, base case model. A comparison of the counterfactual model solution and the true *ex post* equilibrium gives an assessment of the effects of the isolated shock. An equivalent experiment could be undertaken by using the *ex post* equilibrium as the initial equilibrium, and setting the shock of interest to its *ex ante* equilibrium value.

In such analyses, the effects of shocks are not typically additive, so that the effects of a policy shock on the *ex ante* equilibrium differs from the effects of removing it from the *ex post* equilibrium. A double calibrated model can be used to analyze how shocks interact.

An early example of double calibration is found in Hill (1995). Hill employs a simple general equilibrium model in an attempt to isolate the injury to the Canadian economy, measured by employment changes, caused by changes in the world price of imports. Between 1972 and 1980, the Canadian economy was subjected to changing tax rates, factor endowments, preferences and technology as well as world price shocks. While changes to tax rates, factor endowments and world prices were discernable from statistical publications, technological progress proved more elusive. The parameters for technical change were calibrated by Hill so that when all the shocks are introduced to the 1972 economy, the 1980 benchmark data is reproduced as a solution to the system. In the counterfactual simulation, Hill allows all the changes to take place, but fixes the world price of industry output so that the share of domestic goods in domestic consumption remains at 1972 levels. The impact of the trade shocks is obtained from comparing the actual trade shock inclusive data to the counterfactual trade shock free solution.

6.2. Sensitivity analysis

The uncertainties which surround the configuration of exogenous parameter values have also attracted attention. One of the ways in which modellers have responded to such uncertainties is to undertake sensitivity analysis. The overwhelming majority of these analyses focus on the effects of the choice of model elasticities, rather than the values for other model parameters. Modellers have addressed the issue of model sensitivity to elasticity estimates by identifying key model elasticities, and reporting results for alternative elasticity configurations. This approach is termed "limited sensitivity analysis" [Wigle (1991)]. While this procedure can give some sense of whether model results are fragile, it provides no meaningful quantitative measure of robustness.

More rigorous statistical sensitivity analysis procedures have been developed by Wigle (1991) who discusses two classes of systematic elasticity sensitivity analysis used in reporting applied general equilibrium model results. Conditional systematic sensitivity analysis (CSSA) develops a distribution for model results by computing a series of solutions as each elasticity is varied while the others remain constant. Unconditional systematic sensitivity analysis (USSA) computes model results over the entire grid of elasticity configurations. USSA is the most thorough and therefore, the more preferable response to criticisms of elasticity specification, but for most models the computational requirements of such a procedure are prohibitive[8].

Pagan and Shannon (1985) develop an approximation method for performing unlimited systematic sensitivity analysis. Instead of solving the model for each point in the elasticity space explicitly, their procedure analyzes the effects of altering elasticity parameters in a region surrounding the model solution. Because their sensitivity procedure relies on calculations made using a linear approximation of the model solution (which is a function of the elasticity parameters), the computational requirements are considerably less than in unconditional systematic sensitivity analysis, while the procedure retains the flexibility to examine the effects of simultaneous elasticity variations. Pagan and Shannon (1985, 1987) apply their sensitivity analysis to linearized models, while Wigle (1991) demonstrates its use in comparison to conditional and unconditional systematic sensitivity analyses using a levels formulation model.

Other sensitivity procedures in which modellers map *a priori* information about elasticities into the model results have also been developed recently. Harrison and Vinod (1992) and Harrison et al. (1992), develop and apply a global sensitivity analysis procedure in which the model is solved for a sample of elasticities. Their procedure relies on sampling from discrete representations of what are usually continuous elasticity probability density functions. DeVuyst and Preckel (1997) argue that the proposed methodology of Harrison and Vinod introduces an identifiable source of bias into the sampling procedure and propose an alternative way of finding discrete approximations to the continuous pdfs, based on Gaussian quadrature. In both approaches, the model results are weighted by the probability of each elasticity configuration used in their derivation. Repeated sampling allows the modellers to build expected values for the model results.

[8] Wigle (1991) calculates that a USSA using 5 values for each elasticity in an 18 elasticity parameter model would require more than 3 trillion model solutions.

These sensitivity analysis procedures, which have been developed for a model's exogenously specified elasticities, all systematically perturb individual parameter values. Such an approach cannot be applied to the model's calibrated parameters since they are jointly determined from a microconsistent data set – perturbing one calibrated parameter would require the others to readjust to maintain equilibrium, but no unique readjustment exists. Dawkins (1997) develops a sensitivity analysis procedure for the joint set of calibrated parameters by holding the data adjustment process constant and systematically perturbing the unadjusted data from which the calibrated parameters are derived.

6.3. Preadjusting data

Another area of calibration which is coming under scrutiny is the role of data adjustments. Implementing the exact calibration procedures used in micro models requires that the input data be consistent with an initial model equilibrium. However, the basic data which modellers usually rely on does not meet these consistency requirements and modellers undertake adjustments so that it does. The procedures which modellers use are largely *ad hoc* and seldom well documented, but recent work by Dawkins (1998) suggests that the choice of adjustment procedure can affect the statistical properties of the model results.

These data adjustments involve two intertwined processes. The first is selecting single values of each required data point for model calibration. Decisions on these are undertaken when data is collected, and include the choice of one data source over another, the approximation of a desired classification with one found in the data, and the method of aggregation. This process is typically model and data specific. The second is one of deriving consistent data from these point estimates – of ensuring that the data meet the equilibrium conditions of the model. Deriving consistent data sets involves adjusting data starting from an initial estimate. This second process can be executed in a systematic way.

Although no formalized statement of these procedures exists in the literature, the reconciliation of initial data estimates into a benchmark data set for large applied models is typically undertaken in two stages. The first finds consistent values for aggregates: consumption, intermediate demands, and production. At this stage, matrix biproportionality is the paramount restriction on data. So, for example, the total supply of each good in the model must equal the total demand, typically defined as the sum of government consumption, exports, intermediate demand and private domestic consumption. The initial, unadjusted values of these aggregates rely heavily on national accounts data.

The second stage draws on formal adjustment algorithms for balancing a matrix subject to consistency with respect to a set of control totals. The approach is to use the aggregate values derived in the first stage as control totals for the adjustment of submatrices in the model. So, for example, where the model identifies more than one private consumer, the aggregate private domestic consumption for each good can serve as the row control totals for the household consumption submatrix, and the total disposable income by household type can provide the column totals. Similarly, aggregate intermediate demand for a good gives the row totals for the intermediate demand matrix, and total expenditures on intermediate goods by sector (typically found as the residual of total receipts and expenditures on value added) provide the column control totals. Because the control totals are consistent with the biproportionality constraint, the values of the submatrices that are consistent with those control totals also fulfil the biproportionality constraint for the benchmark data set as a whole.

The information required to specify the submatrices in the benchmark data set is typically more detailed than is true for the aggregate values. Initial estimates for the elements of the intermediate demand matrix can be derived from input–output matrices, while those for the household consumption matrix can be derived from household expenditure surveys. Unlike national accounts, such detailed data are unlikely to be collected annually and matrix adjustment is achieved by updating earlier years' estimates so that they are consistent with the benchmark year control values.

Micro modellers do not employ any common approach to data adjustments. They do, however, typically employ *ad hoc* algorithms to derive consistent aggregate values, and resort to formal algorithms, particularly the RAS (Row and Column Scaling) algorithm to derive consistent consumption and production submatrices. RAS is an adjustment algorithm attributed to Bacharach (1970) in which the rows and columns of a matrix are scaled and sequentially updated by the ratio of the matrix row or column sum to the control total row or column sum. This algorithm allows large initial data entries to deviate relatively more from their initial values than small entries.

Other algorithms, most notably those using weighted constrained quadratic minimization, are also employed when making these adjustments. One algorithm, the Stone (1978) and Byron (1978) algorithm is particularly appealing in that it incorporates information about the reliability of the data so that the least reliable data changes more from its initial values than does the more reliable data.

7. Conclusion

In this chapter we discuss rather than debate calibration in its various guises in modern economics. What is it? Why is it used? What is best practice in calibration? How is it evolving?

Calibration, we suggest, is the choice of parameter values subject to a goodness of fit criterion with respect to data, and as such, is conceptually similar to conventional estimation. The common practice has been to apply standards to parameterization which are not used in the econometric literature, such as consistency with a particular base-case, economy-wide model solution for a general equilibrium model, or a long-run, balanced growth path for a dynamic economy. The reason calibrators use these standards, rather than those in more conventional econometric models, is to stay close to particular theoretical models which, in turn, are either hard to estimate, not estimated, or even unestimable. We note that calibration is also a widely used technique in natural and life sciences.

Despite its extensive use, the term calibration is nowhere fully described or defined. We describe the process by discussing in detail some of the calibrations used in the literature, going back to the early 1970s. Not only must the model form be preselected, which can affect results heavily, but typically key parameters, such as elasticities, must be specified. All these decisions affect model results, and consequently, the findings from calibration exercises must be qualified; the interpretation should be suggestive, rather than definitive. Results are conditional on the model structure, the choice of functional form, and key exogenous parameters.

Calibrated models have become a mainstream form of empirical investigation in macroeconomics in recent years, and we also set out how calibration is evolving and changing as a technique in modern economics, drawing contrasts between calibration in macro and micro models. Best practice, we suggest, involves attuning model calibration to questions asked, the choice of appropriate functional forms, and sensitivity analysis of results. New developments in the area focus on double calibration, data preadjustments, and the use of model-estimation consistent elasticities.

References

Abrego, L., and J. Whalley (1998), “Ambiguity enhancing and reducing calculations and the trade and wages debate”, Mimeo (University of Warwick, UK).

Adams, P., and P. Higgs (1990), “Calibration of applied general equilibrium models from synthetic benchmark equilibrium data sets”, Economic Record 66:110–126.

Andolfatto, D. (1996), “Business cycles and labor-market search”, American Economic Review 86: 112–132.

Attanasio, O., and M. Browning (1995), “Consumption over the life cycle and over the business cycle”, American Economic Review 85:1118–1137.

Attanasio, O., and G. Weber (1993), “Is consumption growth consistent with intertemporal optimization? Evidence from the consumer expenditure survey”, Journal of Political Economy 103:1121–1157.

Bacharach, M. (1970), Biproportional Matrices and Input-Output Change (Cambridge University Press, Cambridge).

Browning, M., L.P. Hansen and J.J. Heckman (1999), “Micro data and general equilibrium models”, in: J. Taylor and M. Woodford, eds., Handbook of Macroeconomics, Vol. 1A (North-Holland, Amsterdam) Chapter 8, pp. 543–633.

Byron, R.P. (1978), “The estimation of large social accounting matrices”, Journal of the Royal Statistical Society A 141:359–367.

Caddy, V. (1976), “Empirical estimation of the elasticity of substitution: a review”, Mimeo (Industries Assistance Commission, Melbourne, Australia).

Canova, F. (1995), “Sensitivity analysis and model evaluation in simulated dynamic general equilibrium economies”, International Economic Review 36:477–501.

Chia, N.-C., S. Wahba and J. Whalley (1992), “A general equilibrium-based social policy model for Côte d’Ivoire”, Poverty and Social Policy Series Paper 2 (The World Bank).

Christiano, L.J., and M. Eichenbaum (1992), “Current real business cycle theories and aggregate labor market fluctuations”, American Economic Review 82:430–450.

Clements, K.W. (1980), “A general equilibrium econometric model of an open economy”, International Economic Review 21:469–488.

Coleman II, W.J. (1996), "Money and output: a test of reverse causation", American Economic Review 86:90–111.

Cooley, T.F., and G.D. Hansen (1989), "The inflation tax in a real business cycle model", American Economic Review 79(4):733–748.

Cooley, T.F., and E.C. Prescott (1995), "Economic growth and business cycles", in: T.F. Cooley, ed., Frontiers of Business Cycle Research (Princeton University Press, Princeton) 1–38.

Cooley, T.F., G.D. Hansen and E.C. Prescott (1995), "Equilibrium business cycles with idle resources and variable capacity utilization", Economic Theory 6(1):35–49.

Cox, D.R., and R. Harris (1985), "Trade liberalization and industrial organization: some estimates for Canada", Journal of Political Economy 93(1):115–145.

Dawkins, C. (1997), "Extended sensitivity analysis for applied general equilibrium models", Warwick Economic Research Papers No. 491.

Dawkins, C. (1998), "Choosing a microconsistency algorithm for applied general equilibrium model data", Mimeo (University of Warwick, UK).

De Jong, D.N., B.F. Ingram and C.H. Whiteman (1996), "A Bayesian approach to calibration", Journal of Business and Economic Statistics 14(1):1–9.

den Haan, W.J., and A. Marcet (1990), "Solving the stochastic growth model by parameterizing expectations", Journal of Business and Economic Statistics 8(1):31–34.

DeVuyst, E.A., and P.V. Preckel (1997), "Sensitivity analysis revisited: a quadrature-based approach", Journal of Policy Modeling 19:175–185.

Dixon, P.B., B.R. Parmenter, J. Sutton and D.P. Vincent (1982), ORANI: A Multi sectoral Model of the Australian Economy (North-Holland, Amsterdam).

Ghez, G.R., and G.S. Becker (1975), The Allocation of Time and Goods over the Life Cycle (Columbia University Press, New York).

Greenwood, J., and G. Huffman (1991), "Tax analysis in a real-business-cycle model: on measuring Harberger triangles and Okun gaps", Journal of Monetary Economics 27:167–190.

Greenwood, J., Z. Hercowitz and P. Krusell (1997), "Long-run implications of investment-specific technological change", American Economic Review 87(3):342–362.

Gregory, A.W., and G.W. Smith (1991), "Calibration as testing: inference in simulated macroeconomic models", Journal of Business and Economic Statistics 9:297–303.

Harberger, A.C. (1962), "The incidence of the corporation income tax", Journal of Political Economy 70:215–240.

Harrison, G.W., and H.D. Vinod (1992), "The sensitivity analysis of applied general equilibrium models: completely randomized factorial sample designs", The Review of Economics and Statistics 74:357–362.

Harrison, G.W., R.C. Jones, L.J. Kimbell and R.M. Wigle (1992), "How robust is applied general equilibrium analysis?", Journal of Policy Modeling 15:99–115.

Hertel, T.W., ed. (1997), Global Trade Analysis, Modeling and Applications (Cambridge University Press, New York).

Hill, R. (1995), "Trade shocks and employment change in Canadian manufacturing industries: an applied general equilibrium approach", International Economic Journal 9:73–88.

Hodrick, R.J., and E.C. Prescott (1980), "Post-war US business cycles: an empirical investigation", Working Paper (Carnegie-Mellon University).

Hoover, K.D. (1995), "Facts and artifacts: calibration and the empirical assessment of real business cycle models", Oxford Economic Papers 45:24–45.

Huggett, M. (1996), "Wealth distribution in life-cycle economies", Journal of Monetary Economics 38(3):469–494.

Johansen, L. (1960), A Multisectoral Study of Economic Growth (North-Holland, Amsterdam).

Jones, L.E., R.E. Manuelli and P.E. Rossi (1993), "Optimal taxation in models of endogenous growth", Journal of Political Economy 101:485–517.

Jorgenson, D.W. (1984), “Econometric methods for applied general equilibrium analysis”, in: H.E. Scarf and J.B. Shoven, eds., Applied General Equilibrium Analysis (Cambridge University Press, Cambridge).

Jorgenson, D.W., D.T. Slesnick and P.J. Wilcoxen (1992), “Carbon taxes and economic welfare”, Brookings Papers on Economic Activity, Microeconomics 1992:393–441.

Judd, K.L. (1991), “A review of recursive methods in economic dynamics”, Journal of Economic Literature 29(1):69–77.

Kehoe, T.J. (1985), “Multiplicity of equilibria and comparative statics”, Quarterly Journal of Economics 10(1):119–147.

Kehoe, T.J., and J. Whalley (1985), “Uniqueness of equilibrium in large-scale numerical general equilibrium models”, Journal of Public Economics 28(2):247–254.

King, R.G., and S.T. Rebelo (1990), “Public policy and economic growth: developing neoclassical implications”, Journal of Political Economy 98(5):S126–S150.

King, R.G., and S.T. Rebelo (1993), “Low frequency filtering and real business cycles”, Journal of Economic Dynamics and Control 17(1–2):207–231.

King, R.G., C.I. Plosser and S.T. Rebelo (1988a), “Production, growth, and business cycles I. The basic neoclassical model”, Journal of Monetary Economics 21:195–232.

King, R.G., C.I. Plosser and S.T. Rebelo (1988b), “Production, growth and business cycles II. New directions”, Journal of Monetary Economics 21:309–341.

Kocherlakota, N.R. (1996), “The equity premium: it’s still a puzzle”, Journal of Economic Literature 34(1):42–71.

Koopmans, T.C., and H. Uzawa (1990), “Constancy and constant differences of price elasticities of demand”, in: J.S. Chipman, D. McFadden and M.K. Richter, eds., Preferences, Uncertainty and Optimality: Essays in Honor of Leonid Hurwicz (Westview Press, Boulder, Oxford).

Kydland, F.E., and E.C. Prescott (1982), “Time to build and aggregate fluctuations”, Econometrica 50(6):1345–1370.

Lawrance, E. (1991), “Poverty and the rate of time preference”, Journal of Political Economy 99:54–77.

Lee, H., and D.W. Roland-Holst (1997), “Trade and the environment”, in: J.F. Francois and K.A. Reinert, eds., Applied Methods for Trade Policy Analysis (Cambridge University Press, New York).

Lucas Jr, R.E. (1987), Models of Business Cycles (Blackwell, Oxford).

Lucas Jr, R.E. (1990), “Supply side economics: an analytical review”, Oxford Economic Papers 42:293–316.

Mansur, A., and J. Whalley (1984), “Numerical specification of applied general equilibrium models: estimation, calibration, and data”, in: H.E. Scarf and J.B. Shoven, eds., Applied General Equilibrium Analysis (Cambridge University Press, Cambridge).

McDougall, R., and R. Tyers (1997), “Developing country expansion and relative wages in industrial countries”, in: T.W. Hertel, ed., Global Trade Analysis, Modeling and Applications (Cambridge University Press, New York).

McGrattan, E.R., R. Rogerson and R. Wright (1997), “An equilibrium model of the business cycle with household production and fiscal policy”, International Economic Review 38:267–290.

McKitrick, R.R. (1995), “The econometric critique of applied general equilibrium modelling: the role of parameter estimation”, Discussion Paper 95/27 (University of British Columbia, Department of Economics).

Mehra, R., and E.C. Prescott (1985), “The equity premium: a puzzle”, Journal of Monetary Economics 15(2):145–161.

Nguyen, T.T., C. Perroni and R.M. Wigle (1996), “Uruguay round impacts on Canada”, Canadian Public Policy 22(4):342–355.

Pagan, A.R., and J.H. Shannon (1985), “Sensitivity analysis for linearized computable general equilibrium models”, in: J. Piggott and J. Whalley, eds., New Developments in Applied General Equilibrium Analysis (Cambridge University Press, Cambridge).

Pagan, A.R., and J.H. Shannon (1987), "How reliable are ORANI conclusions?", Economic Record 63:33–45.

Parente, S.L., and E.C. Prescott (1994), "Barriers to technology adoption and development", Journal of Political Economy 102(2):298–321.

Perroni, C., and T.F. Rutherford (1998), "A comparison of the performance of flexible functional forms for use in applied general equilibrium modelling", Computational Economics 11(3):245–263.

Pesaran, M.H. (1974), "On the general problem of model selection", Review of Economic Studies 41(2):153–171.

Piggott, J., and J. Whalley (1985), U.K. Tax Policy and Applied General Equilibrium Analysis (Cambridge University Press, Cambridge).

Piggott, J., and J. Whalley (1998), "VAT base broadening and self supply", Working Paper (NBER).

Prebisch, R. (1962), "The economic development of Latin America and its principal problems", Economic Bulletin for Latin America 7:1-22. First published in 1950 as an independent booklet by UN ECLA.

Prescott, E.C. (1986), "Theory ahead of business cycle measurement", Research Department Staff Report 102 (Federal Reserve Bank of Minneapolis).

Roberts, B.M. (1994), "Calibration procedure and the robustness of CGE models: simulations with a model for Poland", Economics of Planning 27:189–210.

Rupert, P., R. Rogerson and R. Wright (1995), "Estimating substitution elasticities in household production models", Economic Theory 6:179–193.

Scarf, H.E., and T. Hansen (1973), The Computation of Economic Equilibria (Yale University Press, New Haven).

Shoven, J.B., and J. Whalley (1972), "A general equilibrium calculation of the effects of differential taxation of income from capital in the U.S.", Journal of Public Economics 1:281–322.

Shoven, J.B., and J. Whalley (1992), Applying General Equilibrium (Cambridge University Press, Cambridge).

Sims, C.A. (1996), "Macroeconomics and methodology", Journal of Economic Perspectives 10(1): 105–120.

Singer, H.W. (1950), "The distribution of gains between investing and borrowing countries", American Economic Review Papers and Proceedings 40:473–485.

St. Hilaire, F., and J. Whalley (1983), "A microconsistent equilibrium data set for Canada for use in tax policy analysis", Review of Income and Wealth 29:175–204.

Stern, R.M., J. Francis and B. Schumacher (1976), Price Elasticities in International Trade: An Annotated Bibliography (Macmillan, London, for the Trade Policy Research Centre).

Stokey, N.L., and S.T. Rebelo (1995), "Growth effects of flat-rate taxes", Journal of Political Economy 103(3):519–550.

Stone, R. (1978), "The development of economic data systems", in: G. Pyatt and J. Round, eds., Social Accounting for Development Planning (Cambridge University Press, New York) foreword.

Tsigas, M.E., G.B. Frisvold and B. Kuhn (1997), "Global climate change and agriculture", in: T.W. Hertel, ed., Global Trade Analysis, Modeling and Applications (Cambridge University Press, New York).

Viner, J. (1950), The Customs Union Issue (Carnegie Endowment for International Peace, New York).

Watson, M.W. (1993), "Measures of fit for calibrated models", Journal of Political Economy 101(6): 1011–1041.

Whalley, J. (1985a), Trade Liberalization Among Major World Trading Areas (MIT Press, Cambridge, MA).

Whalley, J. (1985b), "Hidden challenges in recent applied general equilibrium exercises", in: J. Piggott and J. Whalley, eds., New Developments in Applied General Equilibrium Analysis (Cambridge University Press, New York).

Whalley, J., and R.M. Wigle (1990), "The international incidence of carbon taxes", in: R. Dornbusch and J. Poterba, eds., Economic Policy Responses to Global Warming (MIT Press, Cambridge).

Wiese, A.M. (1995), "On the construction of the total accounts from the U.S. national income and product accounts: how sensitive are applied general equilibrium results to initial conditions", Journal of Policy Modeling 17:139–162.

Wigle, R.M. (1991), "The Pagan–Shannon approximation: unconditional systematic sensitivity in minutes", in: J. Piggott and J. Whalley, eds., Applied General Equilibrium (Physica Verlag, Heidelberg).

Yang, Y., W. Martin and K. Yanagishima (1997), "Evaluating the benefits of abolishing the MFA in the Uruguay round package", in: T.W. Hertel, ed., Global Trade Analysis, Modeling and Applications (Cambridge University Press, New York).

Young, L.M., and K.M. Huff (1997), "Free trade in the Pacific rim: on what basis?", in: T.W. Hertel, ed., Global Trade Analysis, Modeling and Applications (Cambridge University Press, New York).

Zeldes, S.P. (1989), "Consumption and liquidity constraints: an empirical investigation", Journal of Political Economy 97:305–346.

This Page Intentionally Left Blank

Chapter 59

MEASUREMENT ERROR IN SURVEY DATA

JOHN BOUND[*]

University of Michigan and NBER

CHARLES BROWN

University of Michigan and NBER

NANCY MATHIOWETZ

University of Maryland

Contents

Abstract 3707
Keywords 3707
1. Introduction 3708
2. The impact of measurement error on parameter estimates 3710
  2.1. Special cases 3712
  2.2. General results – linear model 3715
  2.3. Differential measurement error – an example 3716
  2.4. Bounding parameter estimates 3721
  2.5. Contaminated and corrupted data 3723
  2.6. Measurement error in categorical variables 3724
  2.7. Nonlinear models 3727
3. Correcting for measurement error 3728
  3.1. Instrumental variables in the bivariate linear model 3729
  3.2. Multivariate linear model 3733
  3.3. Nonlinear models 3735
  3.4. The contribution of validation data 3737
4. Approaches to the assessment of measurement error 3740
5. Measurement error and memory: findings from household-based surveys 3743
  5.1. Cognitive processes 3743
  5.2. Social desirability 3745

[*] We are grateful to Joe Altonji, Dan Hamermesh, Jeff Wooldridge, Michael Baker, Gary Solon, Shinichi Sakata, participants at the conference for this volume and especially to Jim Heckman and Ed Leamer for extremely helpful comments on previous versions of this paper and to Mari Ellis for help with the preparation of the manuscript.

Handbook of Econometrics, Volume 5, Edited by J.J. Heckman and E. Leamer

5.3. Essential survey conditions 3746
5.4. Applicability of findings to the measurement of economic phenomena 3747
6. Evidence on measurement error in survey reports of labor-related phenomena 3748
6.1. Earnings 3748
6.1.1. Annual earnings 3748
6.1.2. Monthly, weekly, and hourly earnings 3766
6.2. Transfer program income 3770
6.3. Assets 3779
6.4. Hours worked 3784
6.5. Unemployment 3791
6.5.1. Current labor force status, and transitions to and from unemployment 3792
6.5.2. Retrospective unemployment reports 3799
6.6. Industry and occupation 3802
6.7. Tenure, benefits, union coverage, size of establishment, and training 3805
6.8. Measurement error in household reports of health-related variables 3811
6.8.1. Health care utilization, health insurance, and expenditures 3811
6.8.2. Health conditions and health/functional status 3817
6.9. Education 3823
7. Conclusions 3830
References 3833

Abstract

Economists have devoted increasing attention to the magnitude and consequences of measurement error in their data. Most discussions of measurement error are based on the "classical" assumption that errors in measuring a particular variable are uncorrelated with the true value of that variable, the true values of other variables in the model, and any errors in measuring those variables. In this survey, we focus on both the importance of measurement error in standard survey-based economic variables and on the validity of the classical assumption.

We begin by summarizing the literature on biases due to measurement error, contrasting the classical assumption and the more general case. We then argue that, while standard methods will not eliminate the bias when measurement errors are not classical, one can often use them to obtain bounds on this bias. Validation studies allow us to assess the magnitude of measurement errors in survey data, and the validity of the classical assumption. In principle, they provide an alternative strategy for reducing or eliminating the bias due to measurement error.

We then turn to the work of social psychologists and survey methodologists which identifies the conditions under which measurement error is likely to be important. While there are some important general findings on errors in measuring recall of discrete events, there is less direct guidance on continuous variables such as hourly wages or annual earnings.

Finally, we attempt to summarize the validation literature on specific variables: annual earnings, hourly wages, transfer income, assets, hours worked, unemployment, job characteristics like industry, occupation, and union status, health status, health expenditures, and education. In addition to the magnitude of the errors, we also focus on the validity of the classical assumption. Quite often, we find evidence that errors are negatively correlated with true values.

The usefulness of validation data in telling us about errors in survey measures can be enhanced if validation data is collected for a random portion of major surveys (rather than, as is usually the case, for a separate convenience sample for which validation data could be obtained relatively easily); if users are more actively involved in the design of validation studies; and if micro data from validation studies can be shared with researchers not involved in the original data collection.

Keywords

measurement error, validation studies, survey data, validity, reliability, retrospective recall, income, earnings, health, errors in variables, instrumental variables, attenuation

JEL classification: C10, C33, C42, C81, J00

1. Introduction

Empirical work in economics depends crucially on the use of survey data. The evidence we have, however, makes it clear that survey responses are not perfectly reliable. Even such salient features of an individual's life as years of schooling seem to be reported with some error. While economists have been aware of the errors in survey data for a long time, until recently most empirical studies tended to ignore it altogether. However, perhaps stimulated by increases in the complexity of the models we have been estimating, and in particular, with the increasing use of panel data that can seriously exacerbate the effect of measurement error on our estimates, economists have been paying an increasing amount of attention to measurement error[1].

Most assessments of the consequences of measurement error and methods for correcting the biases it can cause have emphasized models that make strong – and exceedingly convenient – assumptions about the properties of the error. Most frequently, measurement error in a given variable is assumed to be independent of the true level of that and all other variables in the model, measurement error in other variables, and the stochastic disturbance. We will refer to such purely random measurement error as "classical" measurement error. In some applications – such as the case where the error is a sampling error in estimating a population mean – these assumptions can be justified. But in most micro data analyses using survey data, they reflect convenience rather than conviction.

From these assumptions comes much of the conventional wisdom about the effects of measurement error on estimates in linear models: (i) error in the dependent variable neither biases nor renders inconsistent the parameter estimates but simply reduces the efficiency of those estimates; (ii) error in the measurement of an independent variable produces downward-biased (attenuated) and inconsistent parameter estimates of its effect, while inadequately controlling for the confounding effects of this variable on the well measured variables; and (iii) the inclusion of other independent variables that are correlated with the mis-measured independent variable accentuates the downward bias[2].

In fact, these conclusions need to be qualified. The bias introduced by measurement error depends both on the model under consideration (e.g., whether it is linear) and on the joint distribution of the measurement error and all the variables in the model. The

[1] Thus, for example, the volatility of earnings and consumption data have often been attributed measurement error [MaCurdy (1982), Abowd and Card (1987, 1989), Hall and Mishkin (1982), Shapiro (1982)]. On the other hand a variety of authors have rationalized a dramatic drop in the magnitude of coefficient estimates associated with the move to fixed effects models in terms of measurement error in key variables and have used a variety of techniques to undo the presumed damage [Freeman (1984) and Card (1996) follow this kind of strategy when using fixed effect models to estimate union premia, Krueger and Summers (1988) do so when estimating industry premia, and Ashenfelter and Krueger (1994) do so when estimating educational premia].

[2] The notion that fixed effect models tend to seriously accentuate the effect of measurement error on parameter estimates represents an important special case of this last point.

effect of measurement error can range from the simple attenuation described above to situations where (i) real effects are hidden; (ii) observed data exhibit relationships that are not present in the error free data; and (iii) even the signs of the estimated coefficients are reversed.

Standard methods for correcting for measurement error bias, such as instrumental variables estimation, are valid when errors are classical and the underlying model is linear, but not, in general, otherwise. While statisticians and econometricians have been quite clear about the assumptions built into procedures they have developed to correct for measurement error, empirical economists have often relied on such procedures without giving much attention to the plausibility of the assumptions they are explicitly or implicitly making about the nature of measurement error. Not only can standard fixes not solve the underlying problem, they can make things worse!

Twenty years ago, analysts would typically have ignored the possibility that the data they were using was measured with considerable error. Rarely, if ever would such researchers acknowledge, let alone try to justify their tacit assumption that measurement error in the data they were using was negligible. More recently, it has become quite common for analysts to correct for measurement error. However, when doing so, researchers virtually always rely on the assumption that measurement error is of the classical type, usually with no justification at all. If we are to be serious regarding measurement error in our data, we need to understand the relationship between the constructs that enter our models and the measures we use to proxy them. This is a tall order. However, even when this "gold standard" is unattainable it will often be possible to put some kind of plausible bounds on the extent and nature of the measurement error of key variables, and use these bounds to work out bounds for estimated parameters of interest.

In addition to providing some evidence about the magnitude of measurement errors, validation studies that compare survey responses to more accurate data such as payroll records permit one to determine whether measurement errors are indeed uncorrelated with other variables. In principle – though this possibility has been realized only incompletely in practice – validation studies can provide more general information on the relationships among errors in measuring each variable, its true value, and the errors and true values in each of the other variables. The research summarized in this chapter is based on direct observation of the measurement error properties of interview reports for a wide range of economic measures. The evidence provides much information to challenge the conventional wisdom.

One general conclusion from the available validation evidence is that the possibility of non-classical measurement error should be taken much more seriously by those who analyze survey data, both in assessing the likely biases in analyses that take no account of measurement error and in devising procedures that "correct" for such error. A second result is that it is important to be at least as explicit about one's model of the errors in the data as about the relationship among the "true" variables that we seek to estimate. Unless one is comfortable assuming that the classical assumptions apply, arguing informally based on that standard case may be dangerous, and writing out the alternative model that better describes one's data can often give real insight into the biases one faces and the appropriateness of traditional cures. A third finding is that, all too often, validation studies are not as helpful to data analysts as they ought to be. Even for the relatively simple goal of assessing the extent of measurement error in individual variables, the "extent" of the error is often not summarized in ways that are suggested by the simple models that guide our thinking. Too few studies take the next step of relating errors in one variable to true values and errors in other variables. We hope that by contrasting the information that is often provided with that which would be most helpful to analysts, we can increase the contribution made by future validation studies. Given the difficulty of mounting a successful validation effort, maximizing the payoff from such efforts is important. In addition to these general themes, we present a variable-by-variable summary of what is known about the accuracy of survey measures.

We begin by reviewing what is known about the impact of measurement error on parameter estimates in Section 2, and possible corrections for the effect of such error in Section 3. This review is not meant as an exhaustive survey the large statistical literature on this subject, but rather is meant to introduce the reader to various issues and to set the stage for our discussion of the validation studies we review[3]. What summary measures of the errors in survey data would be most valuable to an analyst for deciding how important such errors are for his/her analysis? How appropriate are standard techniques for "correcting" for measurement errors, given what validation studies can tell us about such errors? In Section 4 we briefly discuss the design of validation studies, while Section 5 reviews what is known about the circumstances under which phenomena are likely to be well reported by survey respondents. Finally, Section 6 reviews validation studies across a large range on substantive areas. This review is organized by variable, so readers can concentrate on the variables that are most important in their own research. We offer some conclusions in Section 7.

2. The impact of measurement error on parameter estimates

We start with the presumption that we are interested in using survey data to estimate some parameters of interest. These parameters might be means or medians, as is the case when we are interested in tracking the unemployment rate or median earnings, but will often represent more complicated constructs such as differences in means between groups or regression slope coefficients. Measurement error in survey data will typically introduce biases into such estimates. If what we are interested in is the

[3] Fuller (1987) contains a thorough discussion of the biases measurement error introduces into parameter estimates and on standard methods for correcting such biases within the context of the linear model, when measurement error is random (classical). Carroll, Ruppert and Stefanski (1995) contains a more general discussion of the same issues within the context of non-linear models, with considerable attention to models in with errors are not purely random.

estimate of simple means, then, as long as measurement error is mean 0, it will not bias our estimates. However, as is well known, if we are interested in parameters that depend on relationships between variables, then even mean 0 measurement error will typically bias our estimates.

In what follows we will focus primarily on the impact of measurement on parameter estimates within the context of the linear model[4]. Most of the statistics and econometrics literature on the subject has dealt with this case, presumably because it is in this case that the impact of measurement error on parameter estimates can be well characterized[5]. There is a growing literature focusing on the impact of measurement error on parameter estimates within the context of non-linear models; however it remains unclear the extent to which the intuitions we develop within the context of the linear model remain true within this context (see Sections 2.7 and 3.3 for further discussion of this point).

Assume the true model is

$$y^* = X^*\beta + \epsilon, \tag{1}$$

where y^* and ϵ are both scalars and X^* and β are vectors. We will maintain the assumption that ϵ is uncorrelated with X^*. The motivation for this assumption is largely strategic – we are interested in the impact that measurement error has on our estimates and so focus on the case where our estimates would be unbiased in its absence. Instead of X^* and y^*, we observe X and y, where

$$X = X^* + \mu; \quad y = y^* + \nu. \tag{2}$$

In general, we will not assume μ and ν are uncorrelated with X^*, y^* or ϵ. We will use the term classical measurement error to refer to the case where μ and ν are assumed to be uncorrelated with X^*, y^* or ϵ[6], and the term nondifferential [Carroll, Ruppert and Stefanski (1995)] measurement error (in explanatory variables) to refer to the case where, conditional on X^*, X contains no information about y^*[7] implying that μ is uncorrelated with either y^* or ϵ[8].

[4] The framework we present here derives from Bound et al. (1994).

[5] Fuller's excellent monograph focuses solely on the linear model.

[6] For the more general case (i.e., nonlinear modes), this condition needs to be strengthened to refer to the case where μ and ν are assumed to be independent of X^*, y^* or ϵ.

[7] More technically, measurement error in X^* is referred to a non-differential if the distribution of y^* given X^* and X depends only on X^* (i.e., $f(y^* \mid X^*, X) = f(y^* \mid X^*)$).

[8] A few examples may clarify the kind of contexts in which differential measurement can occur. Kaestner, Joyce and Wehbeh (1996) estimate the effect of maternal drug use on an infant's birth weight. They find that self-reported drug use has a larger estimated effect on birth weight than does drug use as assessed from clinical data. They argue that this is because casual users tend to under-report drug use. Thus, if X^* is a binary measure of drug use based on clinical data and X the self report, and y^* is birth

Measurement error in the above sense can occur for a number of reasons that are worth keeping distinct. Respondents can simply misreport on a measure because, for example, their memory is flawed. Here it is possible to imagine obtaining a perfectly measured and therefore valid measure of the quantity in question. An example of this might be pre-tax wage and salary income (i.e., earnings) for a specific calendar year. Alternatively we may be using X and y to proxy for the theoretical constructs of our economic models. Thus, for example, we might use reported years of educational attainment as a proxy for human capital. In this case, the errors will importantly include the gap between what the survey intended to measure and our theoretical construct. While something of the same statistical apparatus can be used to analyze the impact of either kind of error on parameter estimates, clearly validation data can shed light on only the first kind of error.

In the absence of validation data, the analyst observes only X and y. We will be primarily interested in the effect of measurement error on the consistency of our estimates. For this reason, we will not distinguish between populations and samples. The “least squares estimator” of β is

$$\beta_{yX} = [X'X]^{-1}X'y. \tag{3}$$

2.1. Special cases

We will present a general approach to dealing with measurement errors in X^* and y^* which are correlated with the true X, y and ϵ. Before doing so, however, it is useful to highlight a few results that can be derived from the general approach for the biases due to measurement errors when convenient assumptions hold. To simplify discussion of the various biases, we assume throughout that the X’s have been defined so that $\beta_{yX} \geqslant 0$. Consider three special cases.

First, if there is classical measurement error in only one independent variable x_j, the proportional bias in estimating β_j depends on the noise to total variance ratio, $\sigma^2_{\mu_j}/\sigma^2_{x_j}$. In particular, with only one independent variable in the regression, the proportional bias is just equal to this ratio,

$$\beta_{yx_j} = \beta \left[1 - \frac{\sigma^2_{u_j}}{\sigma^2_{x^*_j} + \sigma^2_{u_j}} \right]. \tag{4}$$

weight, $E(y^* \mid X^*, X)$ is decreasing in X. It is also plausible for measurement error to be differential in the context in which X does not merely represent a mismeasured version of X^*, but is a separate variable representing a proxy for X^* [Carroll, Ruppert and Stefanski (1995)]. Thus, for example, if there are contextual effects, the use of aggregate proxies for micro level constructs can exaggerate causal effects [Loeb and Bound (1996), Geronimus, Bound and Neidert (1996)].

With other variables in the regression:

$$\beta_{yx_j \cdot Z} = \beta_j \left[1 - \frac{\sigma^2_{u_j}}{\sigma^2_{x_j^*} \left(1 - R^2_{x_j^*, Z^*} \right) + \sigma^2_{\mu_j}} \right], \tag{5}$$

where Z^* represents the elements of X^* other then x_j^* ($X^* = [x_j^* \mid Z^*]$), $R^2_{x_j^*, Z^*}$ represents the R^2 from the regression of x_j^* on the remaining elements of X^*, and $b_{yx_j, Z}$ represents the least squares regression of y on x_j holding Z constant. Thus, classical measurement error in just one explanatory variable attenuates estimates of the effect of this variable on outcomes. The magnitude of this attenuation depends both on the noise to signal ratio and on the extent of multi- collinearity between the error ridden variable and the other variables in the equation [Levi (1973), Garber and Klepper (1980)].

The measurement error in x_j biases not just estimates of β_j, but also the coefficients on the accurately measured variables. Letting Π represent the coefficient vector from the least squares regression of x_j^* on Z^*, then

$$\beta_{yZ \cdot x_j} = \beta_{i \neq j} + \left(\beta_j - \beta_{yx_j \cdot Z} \right) \Pi. \tag{6}$$

Thus, classical measurement error in x_j^* implies that using x_j as a proxy for x_j^* will partially, but only partially, control for the confounding effects of x_j^* on the estimates of the effect of other variables on outcomes [McCallum (1972), Wickens (1972), Garber and Klepper (1980)].

Second, even if the error, μ_j, is correlated with the true x_j^* (or other X^*'s), but is uncorrelated with ϵ, the proportional downward bias is equal to the regression coefficient from a hypothetical regression of μ_j on the set of measured X's. If there is only one independent variable in the model, this reduces to the simple regression coefficient $\beta_{\mu X}$,

$$\beta_{yx} = \beta \left[1 - \beta_{ux} \right].$$

When μ and X^* are uncorrelated, $\beta_{\mu X}$ is equal to the variance ratio $\sigma^2_\mu / (\sigma^2_{X^*} + \sigma^2_\mu)$ and, as such, will be between 0 and 1. More generally, this will not be true. In particular if μ and X are negatively correlated (the error μ is "mean reverting"), $\beta_{\mu X}$ will typically be smaller than in the classical case (this happens as long as $\sigma^2_\mu < \sigma^2_{X^*}$) and can even be negative – that is, $\beta_{yX} > \beta$. More generally, if the error in one variable is correlated with other variables in the model, the biases on various coefficients depend on the direction of the partial correlation between the error and the various variables in the model.

Third, if the dependent variable y is measured with error, and that error is correlated with the true y^* (where $v = \delta y + v^*$ and v^* is uncorrelated with X^* and ϵ), and the X^*'s are measured without error, then the proportional bias in estimating β is just equal to δ. To emphasize the similarity to the previous case, note that δ is just the regression coefficient β_{vy}.

Each of the above results applies to cross-section analysis, and to panel data by substituting ΔX^* for X^*, etc. But when one uses Δy and ΔX as one's dependent and independent variables, respectively, another aspect of the data becomes important – the correlation over time in the true values (the correlation between y at time t and at time $t-1$, and similarly for X) and in the measurement errors (the correlation between v at time t and at time $t-1$, and similarly for μ). A general result is that, if the variance of a variable (say, X^*) is the same in both years, the variance of ΔX^* is equal to $2\sigma^2_{X^*}(1 - r_{X^*_t, X^*_{t-1}})$ which is greater or less than $\sigma^2_{X^*}$ as $r_{X^*_t, X^*_{t-1}}$ is less than or greater than one-half. A common concern, usually expressed in the context of classical measurement errors, is that true values of X will be highly correlated over time, while the measurement errors will be more or less uncorrelated. In this case, $\sigma^2_{\Delta X^*}$ will be less than $\sigma^2_{X^*}$, while $\sigma^2_{\Delta\mu}$ will be greater than σ^2_{μ}, so that moving from "levels" to "changes" intensifies the bias due to errors in measuring the independent variable(s)[9].

There is one more special case worth noting. Suppose that x_j represents a component of x^*_j[10], with $r_{\mu_j, x_j} = 0$, and that other variables (both y^* and the other X^*'s) are measured without error. Take first the case where x^*_j represents the only explanatory variable in the model. Equation (1) can now be rewritten:

$$\begin{aligned} y &= \beta\left[x_j + \mu_j\right] + \epsilon \\ &= \beta x_j + \left[\beta\mu_j + \epsilon\right]. \end{aligned} \tag{7}$$

Since, by assumption, μ_j is orthogonal to x_j, the composite error in Equation (7) will be orthogonal to x_j and b_{yx_j} will consistently estimate β_j. With other variables in the equation, OLS will no longer consistently estimate β. Rewriting Equation (1) in this case we have:

$$\begin{aligned} y &= \beta_j\left[x_j + \mu_j\right] + Z\gamma + \epsilon \\ &= \beta_j x_j + Z\gamma + \left[\beta_j\mu_j + \epsilon\right]. \end{aligned} \tag{8}$$

While μ_j is orthogonal to x_j, we do not expect μ_j to be orthogonal to Z. Thus, in this case the exclusion of μ_j from our estimating equation represents a specification error, and both $\beta_{yx_j \cdot Z}$ and $\beta_{yZ \cdot x_j}$ will be biased. If the signs of the partial correlations

[9] See Griliches and Hausman (1986) for an illuminating discussion of these issues.

[10] This variance component framework fits many different kinds of contexts. Thus, for example, we might imagine that x^*_j represents schooling, with x_j representing the observed quantity of schooling obtained and μ_j representing the unobserved quality of this schooling (here one might question the orthogonality of μ_j and x_j). Alternatively, x_j might represent cell means of x^*_j (we use industry specific injury rates as proxies for job specific injury rates). Here μ_j and x_j are orthogonal by construction. More detailed discussions of this latter case can be found in Dickens and Ross (1984) and Geronimus, Bound and Neidert (1996).

between μ_j and Z are the same as signs of the partial correlations between x_j and Z, then using x_j as a proxy for x_j^* will only partially control for the confounding effect of Z on β – as an estimate of β, $\beta_{yx_j \cdot Z}$ will be still be biased in the same direction as is β_{yx_j}[11].

2.2. General results – linear model

Having highlighted some special cases in which the consequences of measurement error can be summarized succinctly, we turn to a more general model. With μ and ν potentially correlated with X^* and y^*, the least squares regression coefficient can be rewritten as

$$\begin{aligned}\beta_{yX} &= (X'X)^{-1}X'(X'\beta - \mu\beta + \nu + \epsilon)\\ &= \beta + (X'X)^{-1}X'(-\mu\beta + \nu + \epsilon).\end{aligned} \tag{9}$$

Therefore, the bias of the least squares estimator of β is

$$\beta_{yX} - \beta = (X'X)^{-1}X'(-\mu\beta + \nu + \epsilon). \tag{10}$$

It is useful to collect the measurement errors and their coefficients. Define

$$\gamma \equiv \begin{bmatrix} -\beta \\ 1 \\ 1 \end{bmatrix}, \qquad \omega \equiv [\mu|\nu|\epsilon].$$

Then Equation (10) can be rewritten as

$$\beta_{yX} - \beta = (X'X)^{-1}X'\omega\gamma \equiv A\gamma.$$

If there are k separate variables in the independent-variable matrix X, then A is k by $k+2$. It can be rewritten in a more intuitive form as

$$A = \left[\beta_{\mu X}|\beta_{\nu X}|\beta_{\epsilon X}\right],$$

where the jth column of $\beta_{\mu X}$ consists of the coefficients from regressing μ_j on X, and $\beta_{\nu X}$ and $\beta_{\epsilon X}$ represent the set of coefficients from regressing ν and ϵ on X.

If there is measurement error in only one independent variable X_j^* and if this error is uncorrelated with ϵ, only one column of A will be nonzero, and $A_{jj} = \beta_{\mu_j X_j}$, as claimed in our discussion of special cases. If $\nu = \delta y + \nu^* = \delta X\beta + \delta\epsilon + \nu^*$, and ν^* is uncorrelated with the other variables of the model, and the independent variables

[11] The direction of this bias is easy to work out. If, for example, the partial correlations between each element of Z and x_j are positive, then $\beta_{yZ \cdot x_j} > \gamma$ and $\beta_{yx_j \cdot Z} < \beta_j$.

are measured without error, then $\beta_{\mu X}$ and $\beta_{\epsilon X}$ are a matrix and vector of zeros, and $\beta_{vX} = \delta\beta$. Thus, the proportional bias for each coefficient equals δ.

As the above expression makes clear, with measurement error in more that one explanatory variable, the bias on any particular coefficient will involve multiple terms, and is hard to characterize. What should be clear is that without some knowledge of the distribution of the errors (μ and v), the situation is hopeless – the data put no restrictions on possible values of β.

Even with classical assumptions, measurement error in more than one explanatory variable does not necessarily attenuate the coefficients on the variables measured with error. Theil (1961) derives a useful approximation to the bias in the context of where two variables are measured with error. He imagines we are interested in estimating the relationship:

$$y^* = \beta_1 x_1^* + \beta_2 x_2^* + \epsilon, \tag{11}$$

but observe only error ridden proxies for the x^*'s, x_1 ($x_1 = x_1^* + \mu_1$) and x_2 ($x_2 = x_2^* + \mu_2$). The errors (the μ's) are assumed to be independent of each other, the x's and ϵ and the x^*'s are scaled to have unit variance. Theil shows that when the errors are small

$$\begin{aligned} \beta_{yx_1 \cdot x_2} - \beta_1 &\approx -\frac{\beta_1\lambda_1}{1-\rho^2} + \frac{\beta_2\lambda_2\rho}{1-\rho^2}, \\ \beta_{yx_2 \cdot x_1} - \beta_2 &\approx -\frac{\beta_2\lambda_2}{1-\rho^2} + \frac{\beta_1\lambda_1\rho}{1-\rho^2}, \end{aligned} \tag{12}$$

where ρ represents the correlation between the x^*'s, and the λ's represent the error to total variance ratios for the two variables ($\lambda_j \equiv \sigma^2_{\mu_j}/\sigma^2_{x_i^*}$). Thus, in the multivariate case, the bias on a particular coefficient depends on factors that, as long as ρ is positive, tend to offset each other. In fact, it should be clear that in the two variable case, the bias on the estimated coefficient on the variable measured with less error can be positive[12].

2.3. Differential measurement error – an example

In many cases, assuming that measurement error is classical is a simple (and potentially dangerous) expedient when we have little a priori reason to believe that any other particular assumption would be more plausible. In other situations, however, we have good reason to believe that the errors are differential, and the basis for this belief can help us write down relatively detailed but still manageable models. The growing

[12] When more than one variable is measurement with error, not only is it no longer true that the coefficients on these variables are necessarily attenuated but it is also no longer true that the inclusion of one of the error ridden variables will necessarily reduce the bias on coefficients on accurately measured variables. See Garber and Klepper (1980) for a succinct discussion of these issues.

literature on labor supply of older workers provides a useful example, both because it is relevant for our discussion of survey measures of health and because doing so will allow us to highlight the potential importance of differential measurement error[13].

A large fraction of the men and women who leave the workforce before the age of 62 report health as the reason they do so. Though health is, no doubt, an important determinant of the age at which men and women retire, there are a variety of reasons not to take these self-reports at face value. It seems plausible that men and, to a lesser extent women, rationalize retirement in terms of health even when they retire primarily for other reasons[14]. Myers (1982) has gone so far as to argue that there is no useful information in self-evaluated health. At the same time, for want of alternative measures, econometric analyses of the labor supply decisions of older men and women have generally used respondents' self-assessment of their health. There remain important questions about the validity of self-reported measures of health and therefore of the inferences that can be drawn from studies that use them.

The most common health measures used in retirement research have been global questions such as, "Does health limit the amount or kind of work you can perform?" or "How would you rate your health? Is it excellent, very good, good, fair or poor?" There are a number of reasons to be suspicious of such survey measures [Parsons (1982), Anderson and Burkhauser (1984, 1985), Bound (1991), Waidmann et al. (1995)]. First, respondents are being asked for subjective judgments and there is no reason to expect that these judgments will be entirely comparable across respondents. Second, responses may not be independent of the labor market outcomes we may wish to use them to explain. Third, since health may represent one of the few "legitimate" reasons for a working aged man to be out of work, men out of the labor force may mention health limitations to rationalize their behavior. Lastly, since early retirement benefits are often available only for those deemed incapable of work, men and women will have a financial incentive to identify themselves as disabled, an incentive that will be particularly high for those for whom the relative rewards from continuing to work are low.

Each of these problems will lead to a different kind of bias. The lack of comparability across individuals represents measurement error that is likely to lead to our underestimating the impact of health on labor force participation, while the endogeneity of self-reported health is likely to lead to our exaggerating its impact. Biases in our estimation of health's impact on outcomes will also induce biases on coefficients of any variables correlated with health. Finally the dependence of self-reported health on the economic environment will induce a bias on estimates of the impact of economic variables on participation, regardless of whether we correctly measure the impact of health itself.

[13] The discussion here follows Bound (1991) closely.

[14] Plausibly, this rationalization is not entirely conscious.

As an alternative to using global self-reported health measures, a variety of authors have argued for the use of what have been perceived to be more objective indicators of health: responses to questions about specific health conditions or limitations[15], doctors' reports or information on subsequent mortality[16]. Such proxies are presumed to be more objective than self-reported health measures, though this does not mean that reports of specific conditions are completely reliable (see Section 6.8.2). Moreover, even with perfectly accurate measures of health conditions or mortality, it is not clear that their use as proxies for health give us an accurate indication of the impact of health on labor supply. Part of the problem with "objective" measures is that they measure health rather than work capacity. As long as these health proxies are not perfectly correlated with work capacity – the aspects of health that affect an individual's capacity of work – they will suffer from errors in variables problems. With self-reported health measures we have biases working in opposite directions and, as such, they will have a tendency to cancel each other out. With objective measures there is only one bias, and, as long as the correlation between the proxy and actual health isn't close to perfect, the bias will be quite substantial.

The issues here are important for our understanding not only of the importance of health, but also of the impact of economic variables on early retirement. Both subjective and objective health indicators are correlated with such things as education, race, pre-retirement earnings, and pre-retirement occupation. These factors are also important indicators of early labor market withdrawal. One interpretation of these correlations is that it is those in poor health who leave the workforce before normal retirement age. Alternatively these correlations could be interpreted as reflecting the fact that poor labor market prospects induce men to leave the labor force, but that they then rationalize this behavior by identifying themselves as limited in their ability to work.

The literature that has compared results using a variety of different health measures has tended to find that health seems to play a smaller role and economic variables a greater one when the more objective proxies are used. Most authors have interpreted these results as an indication of the biases inherent in using self-reported measures [Parsons (1982), Anderson and Burkhauser (1984, 1985), Chirikos and Nestel (1981), Lambrinos (1981)]. These authors have typically either ignored the possible biases inherent in the use of a proxy, or have assumed that these biases are small in comparison to the ones introduced by the use of self-reported measures.

Others have argued in favor of using self-reported information [Burtless (1987), Sickles and Taubman (1986)]. These authors emphasize the flaws inherent in most

[15] While responses to questions about specific health conditions or limitations still represent self-reports, the presumption has been that such measures are less susceptible to measurement and endogeneity problems since the questions are narrower and more concrete and, unlike questions about work limitations, are not linked to employment behavior.

[16] For a review of the literature on the effects of health on labor supply decisions see Currie and Madrian (1999).

objective measures of health while pointing to the clinically oriented research supporting the reliability and predictive validity of self-reported health measures [Idler and Benyamini (1997), Nagi (1969), Maddox and Douglas (1973), LaRue et al. (1979), Ferraro (1980), Mossey and Shapiro (1982), Manning et al. (1982)]. These authors ignore the fact that even if self-reported health is a reliable indicator of actual health, this may not be enough to guarantee that it will give sensible results when used as a proxy for health in retirement equations. At issue is whether self-reported health measures are systematically biased, with those out of work being substantially more likely to report health problems than those working. Were this the case, the use of self-reported measures might give misleading information on the reasons why men retire early even if these measures were highly correlated with actual health.

To make these comments precise, we consider a simple model for the labor supply of older men or women. The choice of hours of work, y, depends on the relative rewards of doing so; w, exogenous income (which for simplicity we ignore); unobserved health status, η; and other random components[17], ϵ:

$$y = \beta_1 w + \lambda_1 \eta + \epsilon. \tag{13}$$

We are interested in consistently estimating β_1 and λ_1. We expect β_1 to be positive. Since η is unobserved, the sign of λ_1 is arbitrary, but if larger values of η are associated with better health then we would expect that λ_1 should be positive as well.

We also have an indicator of η, self-reported health h. h depends on health status η, but also on the economic rewards for continuing to work, w, and, again on other random components μ_1,

$$h = \beta_2 w + \lambda_2 \eta + \mu_1. \tag{14}$$

We expect both β_2 and λ_2 to be positive.

We assume that η is orthogonal to both ϵ and μ_1 but, as long as there are common unobserved components that affect both h and y, as there will be if the two are definitionally related or if health limitations act as a rationalization for retirement, ϵ and μ_1 will be positively correlated.

As long as η and w are positively correlated, ignoring η in estimating Equation (13) will lead to overestimates of the importance of economic incentives in determining labor force participation. The obvious alternative would be to use h as a proxy for η but there are a variety of econometric problems with doing so. The correlation between ϵ and μ_1 introduces a simultaneity bias while variance in μ_1 introduces errors-in-variables biases on $\hat{\lambda}_1$. Errors in estimates of λ_1 translate into errors in estimates of

[17] The notational conventions we use in this section are somewhat different than the conventions we use elsewhere. To focus attention on the impact of differential measurement error in our health measure, we are abstracting from potential measurement error in other variables. Thus, the reader should think of y as representing well-measured hours and w as well-measured compensation. On the other hand, h and d represent error ridden measures of health, η.

β_1, while the dependence of h on w introduces an additional bias on $\hat{\beta}_1$. In particular, treating y and h as if they were observable, letting $r_{\eta,w}$ represent the correlation between η and w, and ρ the correlation between ϵ and μ_1 and normalizing λ_2 to equal 1, it is easy to show that:

$$\hat{\lambda}_1 = \frac{\lambda_1 \sigma_\eta^2 (1 - r_{\eta,w}^2) + \sigma_\epsilon \sigma_{\mu_1} \rho}{\sigma_\eta^2 (1 - r_{\eta,w}^2) + \sigma_{\mu_1}^2},$$

$$\hat{\beta}_1 = \beta_1 + \left(\lambda_1 - \hat{\lambda}_1\right) \frac{\sigma_{\eta,w}}{\sigma_w^2} - \hat{\lambda}_1 \beta_2.$$

As long as $\rho > 0$, this correlation will impart an upward bias on $\hat{\lambda}_1$, while $\sigma_{\mu_1}^2$ will impart the standard errors-in-variables downward bias on $\hat{\lambda}_1$. Which one dominates depends on the relative strength of these two forces. The bias on $\hat{\beta}_1$ will depend both on the bias on $\hat{\lambda}_1$ and on β_2. Thus, even if the errors-in-variables and the simultaneity biases on $\hat{\lambda}_1$ were to cancel, we might still tend to underestimate β_1.

The above expressions make clear that the biases on $\hat{\lambda}_1$ and $\hat{\beta}_1$ may be quite substantial even when h is a reliable measure of η (i.e., even when $\sigma_{\mu_1}^2$ is quite small). They also make clear that the magnitude and even the direction of the bias depends on the magnitude of several different correlations. Even if self-reported health is highly correlated with actual health estimates using it as a proxy for health may not give reliable results. Likewise, even if self-reported health often represents rationalization, the use of self-reports may not necessarily exaggerate the role of health in retirement. Beliefs about the kinds of bias involved using self-reported health as a proxy for actual health implicitly reflect judgments about all quantities involved in the above expressions.

Now consider a somewhat more complete model where we have added an equation to make explicit the correlation between w and η and have some more objective indicator of health status, d, which for concreteness sake we will imagine to be subsequent mortality. We have

$$y = \lambda_1 \eta + \beta_1 w + \epsilon, \tag{15}$$

$$h = \lambda_2 \eta + \beta_2 w + \mu_1, \tag{16}$$

$$d = \lambda_3 \nu + \mu_2, \tag{17}$$

$$w = \lambda_4 \eta + \zeta, \tag{18}$$

$$\eta = \nu + \xi. \tag{19}$$

In this model, health η has two components – ν, which influences both longevity and work capacity (e.g., heart problems), and ξ, which influences only the capacity for work (e.g., arthritis). The implicit assumption imbedded in the variance components formulation ($\eta = \nu + \xi$) is that, up to factors of proportionality (λ_1/λ_2 and λ_4/λ_2), ν and

ξ enter the labor force, health and compensation equations with identical coefficients. This assumption seems a natural one as we are thinking of η as capacity for work, and h as a self-report on this capacity. ϵ, μ_1 and μ_2 are assumed to be uncorrelated with w, while all four errors (ϵ, the μ's, and ζ) are assumed to be uncorrelated with η or its components ν and ξ. μ_2 is assumed to be uncorrelated with either ϵ, μ_1 or ζ. These assumptions imply that ζ is also uncorrelated with either ϵ or μ_1. Lastly, ν and ζ are assumed to be uncorrelated with each other. This assumption is mostly definitional – ξ is the piece of η that is uncorrelated with d.

d is objective in two ways that h is not: d does not depend directly on w nor is μ_2 correlated with ϵ. Still, as long as the date of death is not perfectly correlated with an individual's capacity for work, using it as a proxy for health will not adequately control for health, in a regression of y on w (and d). In particular, normalizing λ_3 to equal 1 we have

$$\hat{\lambda}_1 = \lambda_1 \frac{\sigma_\nu^2(1 - r_{\nu,w}^2)}{\sigma_\nu^2(1 - r_{\nu,w}^2) + \sigma_{\mu_2}^2},$$

$$\hat{\beta}_1 = \beta_1 + \left(\lambda_1 - \hat{\lambda}_1\right) \frac{\sigma_{\nu,w}}{\sigma_w^2} - \hat{\lambda}_1 \frac{\sigma_{\xi,w}^2}{\sigma_w^2}.$$

As long as there are disabling conditions that are not life threatening (e.g., severe back problems, mental illness), controlling for d will still leave an omitted variable bias on $\hat{\beta}_1$, while as long as current capacity for work does not perfectly predict date of death there will be errors-in-variables biases on both $\hat{\lambda}_1$ and β_1.

To summarize, using mortality information as a health proxy will tend to underestimate the effects of health and overestimate the effects of economic variables on the labor force participation decision. In contrast, using self-reported health status can either over- or underestimate the impact of either health or economic variables on such decisions.

2.4. Bounding parameter estimates

While, without some restrictions on the nature of the measurement error, the data puts no bounds on β, there has been considerable work done putting bounds on β under the assumption that measurement error is classical. The oldest, and best known of such results is due to Gini (1921). Working with the simple bivariate regression (eqs. 1 and 2, p. 3711) and under the assumption that the errors ν and μ are uncorrelated with each other, with y^* and x^*, and with ϵ, it is easy to show that

$$\frac{1}{\beta_{xy}} = \beta \left[1 + \frac{\sigma_\epsilon^2 + \sigma_\nu^2}{\beta^2 \sigma_{x^*}^2}\right]. \tag{20}$$

Thus

$$\beta_{yx} \leqslant \beta \leqslant \frac{1}{\beta_{xy}}. \tag{21}$$

Under the assumptions of classical measurement error, β_{yx} and $1/\beta_{xy}$ bound β, with the tightness of the bounds being a function of the R^2 between y^* and x^*. More generally, if we allow $r_{x^*,\mu} \neq 0$ but maintain the other assumptions, it is possible to show that as long as the correlation between x and x^* is positive, β_{yx} will be correctly signed [Weinberg, Umbach and Greenland (1994)].

Under the assumption that only one of the explanatory variables is measured with error, it is easy to generalize Gini's result to regressions with multiple explanatory variables. On the other hand, in the context in which multiple explanatory variables are all measured with error, the situation is more complex. Klepper and Leamer (1984) derive results under the assumption that the errors are independent of each other and of the unobserved correctly measured variables[18].

We start by illustrating Klepper and Leamer's result within the context of a model with two explanatory variables,

$$y^* = \beta_1 x_1^* + \beta_2 x_2^* + \epsilon. \tag{22}$$

For ease of discussion, we will assume that x_1^* and x_2^* have been normalized in such a way that β_1 and β_2 are both are non-negative. We can imagine several possible "estimates" of β_1 and β_2. The estimates from the direct regression

$$\hat{\beta}_1^0 = \beta_{yx_1 \cdot x_2}, \quad \hat{\beta}_2^0 = \beta_{yx_1 \cdot x_2},$$

the estimates from the reverse regression of x_1 on y and x_2,

$$\hat{\beta}_1^1 = \frac{1}{\beta_{x_1 y \cdot x_2}}, \quad \hat{\beta}_2^1 = -\frac{\beta_{x_1 x_2 \cdot y}}{\beta_{x_1 y \cdot x_2}},$$

and the estimates from the regression of x_2 on y and x_1,

$$\hat{\beta}_1^2 = -\frac{\beta_{x_2 x_1 \cdot y}}{\beta_{x_2 y \cdot x_1}}, \quad \hat{\beta}_2^2 = \frac{1}{\beta_{x_2 y \cdot x_1}}.$$

These three estimates of β_1 and β_2 represent three points on a two dimensional plane. Klepper and Leamer's results imply that if all three sets of estimates are non-negative, β_1 and β_2 must lie within the triangle defined by these three points and, as a result, $\min[\hat{\beta}_1] \leqslant \beta_1 \leqslant \max[\hat{\beta}_1]$ and $\min[\hat{\beta}_2] \leqslant \beta_2 \leqslant \max[\hat{\beta}_2]$. If, on the other hand, one or more of these estimates is negative, then the first and second moments of the data put no bounds on possible values of β_1 and β_2[19]. Klepper and Leamer show that in

[18] Some of the results developed by Klepper and Leamer had been developed previously by Koopmans (1937), Reiersol (1945), Dhondt (1960), and Patefield (1981), however Klepper and Leamer's treatment of these issues is both the clearest and the most complete.

[19] Klepper and Leamer show that, if all the variables involved are normal, the bounds they derive are tight and that every point within these bounds represents a maximum likelihood estimate of the regression parameters.

the two variable case all possible reverse regression coefficients will be positive if and only if $r_1 r_2 > \rho$, where r_1 and r_2 represent the simple correlations between the two measured explanatory variables and y, and ρ represents the correlation between the two measured explanatory variables. Thus, the higher is the simple correlation between the two explanatory variables and the dependent variable and the lower is the correlation between explanatory variables, the more likely it is that the data will put bounds on the regression parameters.

More generally, in the context where one has k potentially mismeasured explanatory variables, imagine the set of all possible k reverse regressions, one with each of the k potentially mismeasured variables treated as the dependent variable, as well as the usual direct regression. Put these reverse regressions into a common normalized form with the dependent variable on the left-hand side. One now has $k + 1$ vectors of regression estimates for the k mismeasured variables. Klepper and Leamer show that if these $k + 1$ regressions are all positive[20], then their convex hull bounds the true parameters[21].

Krasker and Pratt (1986) take a different approach. In the context of multiple regression where only one of the variables is measured with error, they ask how highly correlated must the error ridden proxy, x_j, be to the unobserved correctly measured variable x_j^* to guarantee that $\beta_{y x_j \cdot Z}$ will be of the correct sign. No assumptions are made about possible correlations between the error (μ_j) and either y^* or any of the elements of X^*. Krasker and Pratt show that as long as

$$r^2_{x_j, x_j^*} > R^2_{x_j^*, Z^*} + \left(1 - R^2_{x_j^*, y^*, Z^*}\right), \tag{23}$$

$\beta_{y x_j \cdot Z}$ will have the correct sign. For the two variable case (where only one is measured with error) they also derive results for $\beta_{yZ \cdot x_j}$. Here, correlations often have to be quite high to guarantee that estimates will be correctly signed.

2.5. Contaminated and corrupted data

The measurement error represented in the typical text book and that has received the most treatment in the statistics literature represents "chronic errors" that affect every observation (the error distributions have no mass point at 0). On the other hand, there are situations in which it may be natural to assume that, while in general a variable is well measured, occasional observations are afflicted with potentially gross errors[22].

[20] Recall that we have normalized X^* in such a way that $\beta \geqslant 0$. More generally, the condition that Klepper and Leamer (1984) derive implies that the data puts bounds on β only if the coefficients from all k possible reverse regressions (regressions of x_j on y and all the other x's) have the pattern of signs as does the original regression of y on X.

[21] Klepper and Leamer also show that, if all the variables involved are normal, every vector of parameter estimates within the convex hull represents a maximum likelihood estimate of the model parameters.

[22] To mention some trivial examples, interviewer errors such as recording that a person was paid 10 dollars per year, rather than per hour, or that person has roughly 10 000, rather than 10 dollars in the bank can lead to occasional gross errors. Imputations for missing data when the researcher is not told which observations include the imputations would be another.

While, formally speaking, our treatment of measurement error in the proceeding sections encompasses this case, intermittent errors are worth some attention on their own.

If one has some notion as to the probability that intermittent errors occur, it is often possible to put bounds on the distribution of the variable of interest. Horowitz and Manski (1995) formalize some quite intuitive ideas. They study the situation in which the researcher is interested in making inference about the marginal distribution of a variable, y_1. However, the researcher does not observe y_1, but rather a random variable y,

$$y = y_1 z + y_0(1 - z),$$

where z represents a random variable that takes on the value of 1 with probability p, 0 with probability $1 - p$; and y_0 a random variable whose distribution is unknown. Horowitz and Mansky refer to the case in which z is independent of y_1 as "contaminated sampling", while the case in which this is not true is referred to as "corrupted sampling"[23].

To see how it is possible to put bounds on the distribution of y_1, imagine we are interested in estimating the median of y_1, m, and suppose we know that $p < 0.1$. It is intuitively clear that, under "contaminated sampling" that m must lie within the closed interval between the 45th and the 55th percentiles of the y distribution (i.e., between the medians of the bottom and top 90% of the y distribution). Under "corrupted sampling", where the missing part of the y_1 distribution could be anything, the bounds are looser. Here, m must lie in the closed interval between the 40th and 60th percentiles of the y distribution.

Horowitz and Manski focus estimating parameters of the marginal distribution of a random variable. It is, however, difficult to apply similar ideas within even the simplest regression context[24]. Thus, for example, if the explanatory variables are potentially "contaminated" in no more than 10% of the sample, one could imagine bounding parameter estimates by throwing out every possible 10% combination of observations. However, with even moderate sample sizes, this procedure would exceed the capacity of current computers.

2.6. Measurement error in categorical variables

While, strictly speaking, the analysis presented in the previous sections applies to both continuous and categorical variables, errors in categorical variables are more

[23] As Horowitz and Manski note, their discussion relates quite closely to discussions within this statistics literature of estimators that are designed to minimize the impact of "contaminated" or "corrupted" data on parameter estimates [Huber (1981), Hampel, Ronchetti, Rousseeuw and Stahel (1986)].

[24] It is possible to see how similar techniques could be used to put bounds on regression coefficients in the context in which the dependent variable suffers from "contaminated" or "corrupted sampling" [Hotz, Mullin and Sanders (1997)]. It is not clear there are practical ways to apply similar ideas in the context in which the items that are "contaminated" or "corrupted" are explanatory variables.

naturally thought of as classification errors. Thus, for example, if x^* is a dichotomous, 0/1 variable, it seems natural to think in terms of the probabilities of false positives ($\pi_{10} \equiv \text{prob}(x = 1 \mid x^* = 0)$) and false negatives ($\pi_{01} \equiv \text{prob}(x = 0 \mid x^* = 1)$). In this context, measurement error cannot be classical. If $x^* = 1$, then $x - x^* \leqslant 0$, while if $x^* = 0$, $x - x^* \geqslant 0$, so it must be the case that $\sigma_{x^*,\mu} < 0$. Thus, errors in binary variables must be mean reverting. More generally, if x^* has a limited range, as is often the case with the constructs we deal with (e.g., educational attainment) there will be a tendency for $\sigma_{x^*,\mu} < 0$ since when x^* is at the maximum (minimum) of its range, reporting errors can only be negative (positive)[25].

Nondifferential measurement error in this context implies that, conditional on the truth, reporting errors are independent of y. In particular,

$$\Pr(x = z_j \mid x^* = z_k, y) = \Pr(x = z_j \mid x^* = z_k), \tag{24}$$

where $z_i \in 0, 1$. This is a strong and often implausible assumption. Suppose, for example, that x represents a chronic health condition – $x^* = 1$ if a person suffers from the chronic condition and is 0 otherwise. It seems plausible that the severity of a person's condition will have an effect on the probability that a person recognizes that they suffer from the condition as well as on outcomes. In this case $\Pr(x = 1 \mid x^* = 1, y)$ will be a function of y, and the random error assumption is violated.

At any rate, under the nondifferential measurement error assumption Aigner (1973) shows that

$$\begin{aligned} \beta_{yx} &= \beta[1 - \Pr(x^* = 1 \mid x = 0) - \Pr(x^* = 0 \mid x = 1)] \\ &= \beta\left[1 - \frac{\pi_{01}\pi}{\pi_{01}\pi + (1 - \pi_{10})(1 - \pi)} - \frac{\pi_{10}(1 - \pi)}{\pi_{10}(1 - \pi) + (1 - \pi_{01})\pi}\right], \end{aligned} \tag{25}$$

where π represents the true fraction of 1's in the population ($\pi = \Pr(x^* = 1)$) and the second line is derived using Bayes rule[26]. Since all the π's lie between 0 and 1, the expression in parenthesis must be less than 1 and β_{yx} will be biased towards 0. In fact, for sufficiently high mis-classification rates (i.e., if $\pi_{01} + \pi_{10} > 1$), β_{yx} can be wrong signed. Bollinger (1996) has worked out bounds for β_{yx} in this model. Under the assumption that $\pi_{01} + \pi_{10} < 1$ and the normalization that $\beta > 0$, Bollinger shows that

$$\beta_{yx} \leqslant \beta \leqslant \max\left(\left[\beta_{xy}\mu_x + \beta_{yx}(1 - \mu_x)\right], \left[\beta_{yx}\mu_x + \beta_{xy}(1 - \mu_x)\right]\right),$$

where $\mu_x \equiv \Pr(x = 1)$. Bollinger also shows how these bounds can be tightened when prior information exists about π_{01} and π_{10}.

[25] As far as we know Siegal and Hodge (1968) were the first to make this point.

[26] If the two kinds of classification error are of the same magnitude (i.e., if $\pi_{01}\pi = \pi_{10}(1 - \pi)$), then the expression in square brackets in Equation (25) simplifies considerably to $1 - \pi_{10} - \pi_{01}$.

Classification error in a dependent variable will also typically bias estimates. Take the case where y^* is a dichotomous, 0/1 variable, and we are interested in estimating $\Pr(y^* = 1 \mid x^*)$. We have accurate measures of x^* $(x = x^*)$ but y suffers from classification error that is independent of x^*, with $\pi_{10} \equiv \Pr(y = 1 \mid y^* = 0)$ and $\pi_{01} \equiv \Pr(y = 0 \mid y^* = 1)$. Since, in this context, the measurement error in the dependent variable is negatively correlated with the accurately measured variable, it should come as no surprise that classification error in a dichotomous dependent variable will tend to bias downward estimates of the effect of y^* on x^*. In fact, it is easy to see that

$$\frac{\partial \Pr(y = 1 \mid x)}{\partial x} = [1 - (\pi_{10} + \pi_{01})] \frac{\partial \Pr(y^* = 1 \mid x^*)}{\partial x^*}. \tag{26}$$

More generally, random misclassification of the dependent variable in a discrete-response setting will bias downwards estimated response functions [Hausman, Abrevaya and Scott-Morton (1998), Abrevaya and Hausman (1997)].

Categorical variables are often thought of as the discrete indicators of continuous latent variables. Thus, we might imagine that $y^* = 1$ if $\xi > 0$ and 0 otherwise. We are interested in estimating $\text{Prob}(y^* = 1 \mid x^*)$, but do not observe y^*. Instead we observe y, where $y = 1$ if $\xi + \nu > 0$ and 0 otherwise. We assume that ν represents normally distributed random "measurement error" in ξ (i.e., ν is assumed to be independent of both ξ and x^*). The probability of classification error in this model depends not just on y^*, but also on ξ and thus on x^*. To keep things simple, we assume that

$$\xi = \beta x^* + \epsilon, \tag{27}$$

where ϵ is a normally distributed, mean 0, random variable. We also assume that x^* is well measured $(x = x^*)$. Were we to directly observe y^*, we could consistently estimate β/σ_ϵ. As it is, however, we can consistently estimate only $\beta/(\sqrt{\sigma_\epsilon{}^2 + \sigma_\nu{}^2})$. Retrieving β/σ_ϵ requires estimated knowledge of σ_ν^2.

Alternatively, imagine that we have a categorical indicator of a latent continuous right hand side variable. Here we imagine the underlying model in terms of the latent variables

$$y^* = \beta\eta + \epsilon. \tag{28}$$

We have a reliable indicator of y^*, y $(y = y^*)$, but observe only a categorical indicator of η, x, where $x = 1$ if $\eta > 0$ and 0 otherwise. In this case, $E(y \mid x = 1) = \beta E(\eta \mid \eta > 0)$, while $E(y \mid x = 0) = \beta E(\eta \mid \eta \leqslant 0)$. Thus, b_{yx} consistently estimates $\beta[E(\eta \mid \eta > 0) - E(\eta \mid \eta \leqslant 0)]$. Now suppose x only imperfectly indicates whether $\eta > 0$. In particular, we assume that $x = 1$ if $\eta + \mu > 0$, where μ represents random measurement error. In this case

$$\beta_{yx} = \beta\left[E(\eta \mid \eta + \mu > 0) - E(\eta \mid \eta + \mu \leqslant 0)\right] < \beta\left[E(\eta \mid \eta > 0) - E(\eta \mid \eta \leqslant 0)\right]. \tag{29}$$

Thus, once again, the use of a noisy explanatory variable tends to lead to an underestimation of the magnitude of the parameter of interest.

2.7. Nonlinear models

While there is growing literature on the impact of measurement error on parameter estimates within the context of non-linear models, discussions universally occur within the context of specific models. For this reason, it is not possible to summarize results in quite the same way as we were when talking about the linear model. Broadly speaking, the results that do exist suggest that (i) results based on linear models are often approximately true within the context of the non-linear models that have been explicitly studied, and (ii) if anything, non-linearities tend exacerbate biases introduced by measurement error.

We have seen that with multiple covariates measured with error, even in the context of the linear model, the effects of measurement error are not easily summarized. On the other hand, in the context of classical measurement error in one variable the bias is always in the form of attenuation. With multiple variables measured with error or if measurement error is not classical, attenuation may not hold.

Weinberg, Umbach and Greenland (1994) study the effect of non-differential measurement error in an explanatory variable within the context of a simple bivariate model, $f(y^* \mid x^*)$, where the "dose-response" is monotonic (i.e., $E(y^* \mid x^*)$ monotonicaly increases (decreases) with x^*). Recall that non-differential measurement error in x^* implies that $f(y^* \mid x^*, x) = f(y^* \mid x^*)$. Weinberg, Umbach and Greenland show that as long as $E(x \mid x^*)$ increases monotonicaly with x^*, σ_{y^*,x^*} and $\sigma_{y^*,x}$ must have the same sign. To paraphrase Weinberg, Umbach and Greenland, as long as the measurement of x^* is good enough that the population mean of measured "exposure" goes up when true "exposure" does, trend reversal can not occur.

While Weinberg, Umabach and Greenland's results suggest that in simple models non-differential measurement error of the kind they describe can not cause trend reversal, monotonicity is not necessarily maintained. Hwang and Stefanski (1994) show that even within the context of classical measurement error, it is possible to find situations where the regression of y^* on x^*, $E(y^* \mid x^*)$, is monotonically increasing (decreasing) in x^*, but that the regression of y^* on x, $E(y^* \mid x)$ is not.

There is also evidence within the context of specific models that non-linearities tend to exacerbate the magnitude of the bias introduced by measurement error. Griliches and Ringstad (1970) analyzed the situation where y^* is a quadratic function of x^*

$$y^* = \beta_0 + \beta_1 x^* + \beta_2 x^{*2} + \epsilon. \tag{30}$$

y^* is assumed to be well measured ($y = y^*$), but x^* is not ($x = x^* + \mu$). Under the assumption that both x^* and μ are normally distributed and that μ is uncorrelated with either x^* or ϵ, Griliches and Ringstad showed that

$$\beta_{yx\cdot x^2} = \beta_1(1-\lambda), \quad \beta_{yx^2\cdot x} = \beta_2(1-\lambda)^2, \tag{31}$$

where, as before, $\lambda = \sigma_\mu^2/\sigma_x^2$. Thus, the coefficient on the quadratic term is more severely biased than is the coefficient on the linear term.

Yatchew and Griliches (1985) derive results for the probit model with one mismeasured explanatory variable. Once again, assuming all variables are distributed normally and that measurement error is classical, they show that simply using x in place of x^* produces estimates that converge to

$$\beta \frac{\sigma_{x^*}^2 / \left(\sigma_{x^*}^2 + \sigma_{\mu}^2\right)}{\sqrt{\sigma_{\epsilon}^2 + \beta^2 \frac{\sigma_{\mu}^2 \sigma_{x^*}^2}{\sigma_{x^*}^2 + \sigma_{\mu}^2}}}. \tag{32}$$

As is evident from Equation (32), the usual bias towards zero that is present in the linear model is compounded by the term appearing after the plus sign in the denominator.

In the linear model, biases due to measurement error do not depend on whether that error is normal or homoskedastic. However, in non-linear models, this is potentially important, and can induce biases that run counter to our intuitions in the linear case. Consider, for example, a Tobit model

$$y^* = x^*\beta + \epsilon, \qquad y = \begin{cases} y^* & \text{if } y^* > 0, \\ 0 & \text{otherwise.} \end{cases}$$

Our explanatory variable x^* is measured with error, and suppose the error μ is heteroskedastic. Then we can re-write the model for the latent variable y^* as

$$y^* = x\beta + (-\mu\beta + \epsilon),$$

where the "error term" in parentheses is heteroskedastic. Given that heteroskedasticity by itself leads to inconsistent parameter estimates in Tobit models, and can in plausible cases lead to over-estimating β [Maddala (1983, p. 179)], it seems quite possible that heteroskedastic measurement error could lead to upward-biased parameter estimates.

3. Correcting for measurement error

Under the assumption that measurement error is classical, statisticians and econometricians have developed a number of methods to deal with the biases introduced into our estimators when measurement error is present. In particular, under such assumptions, knowing the marginal distribution of the u_j's is sufficient to allow the researcher to undo the biases introduced by measurement error. Alternatively, if one has exogenous determinants of the error ridden explanatory variables or, in some cases, multiple indicators of the same outcome, one can use these as instruments[27, 28].

[27] The focus of this section is on point estimation. As such, we ignore sampling variability of the various estimators we discuss. In many cases, the estimators are or can be interpreted as instrumental variable estimators. More generally a discussion of the distribution of these estimators can be found in Fuller (1987), Carroll, Ruppert and Stefanski (1995) and Newey and McFadden (1994).

[28] The methods mentioned all involve introducing external information. As long as the measurement error in X^* is classical, and X^*, itself, is not normally distributed, β is formally identified [Reiersol

We wish to emphasize three points about such general strategies. The first is that these strategies are not as distinct as they might first seem. The second is that these strategies for obtaining consistent estimates of the parameters of interest work if measurement is classical, but do not, in general do so otherwise. Third, even when the correction does not produce consistent estimates, it may produce a bound; and if OLS and IV are biased in different directions or IV is less biased than OLS, this additional information may be very valuable.

3.1. Instrumental variables in the bivariate linear model

To illustrate these points we will focus on the bivariate linear regression model. To further simplify things, we will also assume that all variables are measured as deviations around their respective means. Thus our model becomes

$$y^* = \beta x^* + \epsilon. \tag{33}$$

We assume that we measure y^* without error ($y = y^*$). On the other hand, we have two error ridden indicators of x^*, $x_1 = x^* + \mu_1$ and $x_2 = x^* + \mu_2$, with μ_1 and μ_2 uncorrelated with x^*.

Using either x_1 or x_2 as proxies for x^* will lead to estimates of β that are biased towards 0. One alternative would be to use the multiple measures of x to first gauge the magnitude of the errors and then to correct the bias introduced by these errors. In particular, under the assumptions that μ_1 and μ_2 are uncorrelated with all the other variables in the system (including each other), $\sigma_{x_1,x_2} = \sigma^2_{x^*}$. Define

$$\lambda_1 \equiv \frac{\sigma^2_{x^*}}{\sigma^2_{x^*} + \sigma^2_{\mu_1}} = \frac{\sigma_{x_1,x_2}}{\sigma^2_{x_1}}, \tag{34}$$

where λ_1 represents the signal to total variance ratio for x_1. Similarly,

$$\lambda_2 \equiv \frac{\sigma^2_{x^*}}{\sigma^2_{x^*} + \sigma^2_{\mu_2}} = \frac{\sigma_{x_1,x_2}}{\sigma^2_{x_2}}. \tag{35}$$

$\beta_{yx_1} = \lambda_1\beta$ and $\beta_{yx_2} = \lambda_2\beta$. Under the assumptions of the model, data on y, x_1 and x_2 allow one to consistently estimate β_{yx_1}, β_{yx_2}, λ_1 and λ_2 and thence β. In fact, two such

(1950), Kapteyn and Wansbeek (1983)]. Under the assumption that X^* is not normal a number of authors have suggested instrumental variable estimators that use third or higher moments of the various variables as instruments for X [Geary (1942), Pal (1980), Cragg (1997), Dagenais and Dagenais (1997), Lewbel (1997)]. However, these methods depend crucially on the assumption that $E(y^* \mid X^*)$ is a strict linear function of X^*, and, as such, estimates will be sensitive to specification error. At any rate, such methods have seldom been used in practice. Alternatively, Wald (1940) suggested an estimator of β that involved grouping the data. However, unless one has some external information that can be used to form groups (i.e., an instrument), the resulting estimator will typically be no less biased than OLS [Pakes (1982)].

estimates are available, giving us some capacity to test the underlying assumptions of the model. In particular, our assumption that μ_1 and μ_2 are uncorrelated with x^* and ϵ implies $\sigma_{x_1,y} = \sigma_{x_2,y}$, which is testable.

Alternatively, one might choose to use x_2 to instrument x_1[29]

$$\beta^1_{iv} = \frac{\sigma_{y,x_2}}{\sigma_{x_1,x_2}}. \tag{36}$$

Notice that $\beta^1_{iv} \equiv \beta_{yx_2}/\lambda_2$. Thus, using x_2 to instrument x_1 is equivalent to regressing y on x_2 and then using an estimate of λ_2 to disattenuate the resulting estimate of β[30].

Under what circumstances will β^1_{iv} represent a consistent estimate of β? To see, we first write out β^1_{iv} in terms of the x^*

$$\beta^1_{iv} = \frac{\beta\left[\sigma^2_{x^*} + \sigma_{x^*,\mu_2}\right] + \sigma_{\mu_2,\epsilon}}{\left[\sigma^2_{x^*} + \sigma_{x^*,\mu_2}\right] + \sigma_{x^*,\mu_1} + \sigma_{\mu_1,\mu_2}}. \tag{37}$$

Thus, $\beta^1_{iv} = \beta$ if $\sigma_{\mu_2,\epsilon} = \sigma_{x^*,\mu_1} = \sigma_{\mu_1,\mu_2} = 0$. In other words $\beta^1_{iv} = \beta$ if x_2 is exogenous, the measurement in x_1, μ_1, the measurement error in x_1, is uncorrelated with x^*, and the measurement errors in x_1 and in x_2 are uncorrelated with each other. These are clearly strong assumptions[31].

The assumption that $\sigma_{\mu_2,\epsilon} = 0$ means that reporting errors in x_2 are unrelated to factors other than the x^* affecting y. There are circumstances where this assumption may be a sensible one, but others in which it is clearly not. For example, if x_1 and x_2 represent two self-reported measures of health, and y represents a measure of labor supply, we might expect that reporting (the μ's) would be correlated with the equation error (ϵ).

The assumption that the two errors in reporting x^* are uncorrelated (i.e., that $\sigma_{\mu_1,\mu_2} = 0$) will also often be open to question. For example, if x_1 and x_2 represent two reports on x^* taken from the same individual but at different times, it seems likely that the two errors will be positively correlated. Even if x_1 and x_2 represent two reports on x^* taken from different individuals it will often be possible that the errors will be positively correlated. Thus, for example, two siblings' reports on their

[29] Of course, instruments don't necessarily have to be alternative indicators of x^*. Any variable w, such that $\sigma_{x^*,w} \neq 0, \sigma_{w,\epsilon} = 0$, and $\sigma_{w,\mu} = 0$ represents a valid instrument for x.

[30] We have been talking as if y, x_1 and x_2 all come from the same sample, but what is often the case is that a researcher has only one measure of x^* in the primary data set of interest, but has an estimate of λ from some other data set which included multiple measures of x. Using an estimate of λ based on one sample to correct regression estimates from another is fine as long as one can justifiably interpret the samples as representing similar samples from similar populations.

[31] Fuller (1987) states these conditions somewhat differently. Using x_2 to instrument x_1 will consistently estimate β if (1) $\sigma_{x_2,\epsilon} = 0$ and (2) $\sigma_{x_2,\mu_1} = 0$. $\sigma_{x^*,\mu_1} = \sigma_{\mu_1,\mu_2} = 0 \Rightarrow \sigma_{x_2,\mu_1} = 0$. Since $\sigma_{x^*,\mu_1} = 0$ and $\sigma_{\mu_1,\mu_2} = 0$ represent conceptually distinct conditions, we think it makes sense to distinguish the two when discussing conditions for the consistency of the IV estimator.

parent's education will usually both be based on what that parent told the two. If the parent exaggerates her educational attainment (e.g., claims to have finished college, even though she did not), it seems likely that this exaggeration will be common to both siblings' reports as well as to the parent's. Moreover, if part of the problem is not simply that individuals inaccurately report x, but that our measures do not accurately reflect our constructs (we are interested in human capital, but ask about educational attainment in years), once again it seems likely that the errors from the separate reports will be positively correlated. In all these situations we expect $\sigma_{\mu_1,\mu_2} > 0$.

Thus, it seems likely that in many situations reporting errors will be positively correlated with each other. The good news here is that, as long as it is true that $\sigma_{\mu_2,\epsilon} = \sigma_{x^*,\mu_1} = 0$, then $\beta \geqslant \beta_{iv}^1 \geqslant \beta_{yx_1}$. Thus, correcting for measurement error will tighten our bounds on the true parameter. In addition, with more than two measures of x^* it is possible to begin to relax some, but not all of the assumptions regarding the independence of reporting errors.

Finally, what about the assumption that $\sigma_{x^*,\mu_1} = 0$? This assumption is really at the heart of classical measurement error model. There are situations where this assumption seems quite reasonable. Thus, for example, if x represents a sample mean and x^* a population mean, then there may be good reason to believe that $\mu = x - x^*$ is independent of x^*. Alternatively, if x^* represents IQ and x the performance on a specific test, then again it may be natural to assume that μ (testing error) is uncorrelated with the truth (here one might want to claim that this is true by construction). However, in the context of survey measurements, there does not seem to be any compelling reason to believe that measurement error is uncorrelated with the truth. Moreover, there are a number of circumstances where it seems likely that reporting errors are negatively correlated with the truth $\sigma_{x^*,\mu} < 0$. For example, if, as may often be the case, x represents a component of x^*, it may be as natural to assume that μ and x are uncorrelated as it does that μ and x^* are uncorrelated. Of course, $\sigma_{x,\mu} = 0$ implies that $\sigma_{x^*,\mu} < 0$.

As we have already mentioned, if $\sigma_{x^*,\mu_1} < 0$, then it is no longer necessarily the case that $\beta_{yx_1} < \beta$. If it is still true that $\sigma_{\mu_2,\epsilon} = \sigma_{\mu_1,\mu_2} = 0$, then $\beta_{iv}^1 \geqslant \beta$. More generally, if all we know is that $\sigma_{x^*,\mu_1} \neq 0$, then β_{yx_1} could either over or under estimate β and exactly the same could be said for β_{iv}^1. Short of some clear notions regarding the nature of measurement error, it is unclear whether standard methods of correcting for biases introduced into our estimates by such errors take us any closer to the truth.

An interesting example of the situation where $\sigma_{x^*,\mu_1} < 0$ occurs in the situation discussed above where x^* is dichotomous, and errors are therefore errors of classification. Now suppose one has two indicators of x^* available, x_1 and x_2[32]. We

[32] There are situations in which the researcher knows or has estimates of π_{01} and π_{10} from external information. Thus, for example, researchers studying the impact of training programs on the employment and earnings of those trained sometimes do not have an explicit control group, but use nationally representative samples instead. In this context, the control group sample will be "contaminated" with

assume that, conditional on x^*, the two measures are independent of each other and of y. In particular this implies

$$\Pr\left(x_1 = z_j \mid x^* = z_k, y, x_2\right) = \Pr\left(x_1 = z_j \mid x^* = z_k\right), \tag{38}$$

and

$$\Pr\left(x_2 = z_j \mid x^* = z_k, y, x_1\right) = \Pr\left(x_2 = z_j \mid x^* = z_k\right). \tag{39}$$

In other words, we are assuming the measurement error in x^* is nondifferential. Here, one might be tempted to use x_2 to instrument x_1; however, as our discussion above will have made clear, this procedure will tend to produce estimates of β that are too large in magnitude. In fact, it is easy to show that

$$\beta_{iv}^1 = \beta \frac{1}{1 - (\pi_{01} + \pi_{10})}, \tag{40}$$

which will be greater than β as long as there is any measurement error in x_1.

However, under the specified assumptions, it is possible to derive consistent estimates of β using GMM methods [Kane, Rouse and Staiger (1999) and Black, Berger and Scott (2000) also mention this possibility]. To see the plausibility that this is the case, it is sufficient to count parameters and moments. The "structural" model includes three parameters: the constant term, the slope coefficient and the error variance. In addition, there are four distinct error rates as well as the probability that $x^* = 1$, a total of eight parameters in all. With data on y, x_1 and x_2 we have 8 independent moments. The cross tabulation of x_1 and x_2 give us three, the mean of y conditional on x_1 and x_2 gives us four more, and the variance of y gives us one – eight in all[33].

More generally, if one is working with a linear model that includes categorical variables and if one has multiple, error-ridden indicators of such variables where the

individuals who received training. However, in these situations, the researcher will typically have reliable information on the fraction of the population that receives training, and can use this as an estimate of π_{01}. At any rate, in this kind of situation it is reasonably straightforward to derive consistent estimators of the parameters of interest. For a discussion of the case where misclassification occurs in an explanatory variable, see Aigner (1973), Freeman (1984), Heckman and Robb (1985) and Heckman, LaLonde and Smith (1999). For the case where the misclassification occurs in the dependent variable, see Poterba and Summers (1986, 1995).

[33] Kane, Staiger and Rouse's work echoes earlier work of Goodman (1974a,b), Haberman (1977), Andersen (1982) and others on what Goodman refers to as latent structural models. Goodman showed that in a context in which one observed multiple independent discrete indicators of a (latent) discrete random variable it was often possible to identify the distribution of both the underlying latent variable and the transition matrices that stochastically map the latent variable into observable indicators. The correspondence between latent structural models and the model proposed by Kane, Rouse and Staiger is remarkably close. However, the models that Goodman and his colleagues worked with involve solely discrete variables and have been mostly ignored by economists.

errors are independent of either the outcome or the other explanatory variables in the system, it is possible to get consistent estimates of the parameter of the model using GMM techniques [Kane, Rouse and Staiger (1999)][34].

While the assumption that $\sigma_{\mu_2,\epsilon} = \sigma_{x^*,\mu_1} = \sigma_{\mu_1,\mu_2} = 0$ is sufficient to identify β, it is not sufficient to fully identify the model. Counting sample covariances makes this clear. $\mathrm{Var}(y, x_1, x_2)$ contains a total of 6 separate terms. However, even with the stated restrictions, our model includes seven distinct parameters (β, $\sigma^2_{x^*}$, σ^2_ϵ, $\sigma^2_{\mu_1}$, $\sigma^2_{\mu_2}$, σ_{x^*,μ_1} and $\sigma_{\mu_i,\epsilon}$). In particular, the conditions necessary for the consistent estimation of β are not sufficient to allow us to separately identify $\sigma^2_{x^*}$, σ^2_ϵ, $\sigma^2_{\mu_1}$ and $\sigma_{\mu_i,\epsilon}$. The IV estimator allows us to solve both the pure errors in variable and the endogeneity problems associated with the use of x_1 as a proxy for x^*, but does not allow us to separate out these two effects. If, in addition to the assumptions we have already made, we assume that μ_2 is uncorrelated with x^* ($\sigma_{x^*,\mu_2} = 0$), or that μ_1 is uncorrelated with ϵ ($\sigma_{\mu_1,\epsilon} = 0$) then the model is fully identified.

As Goldberger (1972) and Griliches (1974, 1986) have emphasized, it is often also possible to consistently estimate errors in variables models in a multi equation setting. We illustrate with an extremely simple model. Suppose

$$\begin{aligned} y_1 &= \beta_1 x^* + \epsilon_1, \\ y_2 &= \beta_2 x^* + \epsilon_2, \\ x &= x^* + \mu. \end{aligned} \tag{41}$$

The error terms (the ϵ's and μ) are assumed to be uncorrelated with each other and with x^*. Under these assumptions, β_1 can be consistently estimated by using y_2 as an instrument for x in the regression of y_1 on x ($\beta_{iv} = \mathrm{Cov}(y_1, y_2)/\mathrm{Cov}(x, y_2)$). β_2 can be estimated in a similar fashion. Chamberlain and Griliches (1975) used more sophisticated multi-equation models to control for "ability" when estimating the effect of education on earnings. However, as Griliches has emphasized, estimates based on such models are only as good as the models themselves. In this kind of setting, minor specification errors can have significant effects on parameter estimates. Griliches (1986) and Aigner et al. (1984) include excellent discussions of these kind of models.

3.2. Multivariate linear model

The methods we have been discussing generalize to the multivariate case. Suppose, for example, one is willing to assume that errors in both the outcome and the explanatory

[34] It is worth noting that the discussion has been of models in which x^* is, itself, categorical. Such models need to be distinguished from models in which x^* is conceptualized as continuous (e.g., health status), but we have only categorical indicators of x^*. If x represents an error ridden categorical indicator of x^* (i.e., if $x = k$ iff $c_{k-1} < x^* + \mu \leqslant c_k$) there may be no particular reason to believe that μ is correlated with x^*. In fact, in this case, the models are linear in latent variables. For this reason, the intuitions and insights obtained from work on the linear errors in variables model still holds. The case where x represents a categorical indicator of an underlying continuous variable has been extensively analyzed [e.g., Heckman (1978), Lee (1982a,b), Muthen (1983)].

variables (ν and μ) are uncorrelated with both the actual (accurately measured) outcome and the explanatory variables and that one has prior knowledge of their joint distribution, then

$$\hat{\beta} = \left(S_{XX} - \hat{\Sigma}_{\mu,\mu}\right)^{-1} \left(S_{Xy} - \hat{\Sigma}_{\mu,\nu}\right), \tag{42}$$

will consistently estimate β, where S_{XX} represents the sample variance of X, S_{Xy} the sample covariance of X and y, $\hat{\Sigma}_{\mu,\mu}$ a consistent estimate of the variance of the μ's and $\hat{\Sigma}_{\mu,\nu}$ a consistent estimate of the covariance between μ and ν.

Alternatively, if one has as many instruments (W's) as one has as one has explanatory variables (X's)[35], ν is uncorrelated with y^*, and $\sigma_{W,\mu} = \sigma_{W,\nu} = \sigma_{W,\epsilon} = 0$ then the IV estimator,

$$\beta_{\mathrm{IV}} = \left[W'X\right]^{-1} W'y, \tag{43}$$

consistently estimates β. Of course, if the assumptions are violated and W is correlated with μ, ν or ϵ, b_{iv} will be inconsistent. One special case is worth noting. Take the situation where only one element of X is measured with error (denote this variable as x) while the rest are accurately measured (denote this vector as Z^*). We are interested in estimating the equation

$$y^* = \beta x^* + Z^{*\prime}\gamma + \epsilon. \tag{44}$$

We have a proxy for x^*, x ($x = x^* + \mu$), but accurately observe Z^* ($Z = Z^*$). We also have available factors that help predict x^*, w. w is uncorrelated with μ, ν, or ϵ. To estimate (β, γ) we use $W = w : Z$ as an instrument for $x : Z$. Under these assumptions, the IV estimator will consistently estimate β, but will consistently estimate γ only if Z is also uncorrelated with μ. Thus, if reporting behavior depends not just on x^*, but also on Z, then the instrumental variables estimator will not consistently estimate γ.

Precisely this kind of situation arises within the context of the example discussed at some length in Section 2.3 above where we were interested in estimating the effect of health and financial factors on retirement behavior. Here x^* represents overall health, x a self-reported indicator of overall health, and Z^* represents other factors including financial ones that effect retirement behavior. As discussed above, it is natural in this context to imagine that measurement error in x^* will be differential – poor health will be used to rationalize behavior. Compared with the global measures, more detailed health indicators available in some surveys (e.g., the Health and Retirement Survey) such as reports of specific chronic conditions or functional limitations may be less susceptible to measurement and endogeneity problems, since the questions are narrower and more concrete. However, as long as such measures

[35] Accurately measured x's can be included as elements of W.

represent only a component of health using such measures directly in labor supply equations is, for the reasons discussed above [see also Bound, Schoenbaum and Waidmann (1995)], likely to lead researchers to underestimate the effect of health and overestimate the effect of financial incentives on retirement behavior. As an alternative to either using the global or detailed health measures in estimating equations, some researchers [Stern (1989), Bound et al. (1999)] have used detailed measures as instruments. However, in this context it would seem natural to worry about the possibility that the rewards for continued work would influence reporting behavior (e.g., those with low rewards for continued work might be particularly likely to report themselves in poor health to justify labor force exit) – x depends not just on x^*, but also on Z^*. In this context, using some exogenous determinants of health along with Z to instrument $x : Z$ will consistently estimate β, but not γ[36].

3.3. Nonlinear models

Correcting for the bias created by errors in variables is more difficult in non-linear than in linear models. Typically, instrumental variable methods work well only when errors are relatively small in magnitude [Amemiya (1985, 1990)]. Thus, for example, suppose one is interested in estimating the non-linear model,

$$y = g(x^*;\ \theta) + \epsilon, \tag{45}$$

where we assume that ϵ is independent of x^*, and that θ is a parameter vector. We observe a proxy for x^*, x, where $\mu = x - x^*$ is independent of x^*. We also have available instruments, w, that are correlated with x^*, but are independent of both μ and ϵ. We

[36] Following the example of Section 2.3 in detail, with two indicators of η we might be tempted to use one to instrument the other, but this will not work. As long as $\beta_2 \neq 0$ using d^* to instrument h^* will purge h^* of its dependence on ϵ and so will correctly estimate λ_1 but will tend to underestimate β_1 by $\beta_2\lambda_1$. The intuition that we should be able to use d^* to instrument h^* arises from the similarity of this model to the classical errors-in-variables model, in which one error-prone measure can be used to instrument another. This model differs from the classical errors-in-variables model in that the endogeneity of h^* causes the error in this indicator to be correlated with the other regressor in the model, w. The instrumental variable procedure uses the projection of h^* onto w and d^* as a proxy for η. What we need, instead, is the projection of η on w and d^*. With h^* as the dependent variable, the estimated coefficient on w will reflect not only the errors in d^* but also w's direct effect on h^*, β_2. This, in turn, will induce the downward bias on β_1 of $\beta_2\lambda_1$. We could sort all of this out if we had a consistent estimate of β_2, but this requires either knowledge of the reliability of d^* as a proxy for η or another indicator of η. Thus, using mortality information to instrument self-reported disability status will correctly estimate the impact of health but tend to underestimate the impact of economic variables on such decisions. In contrast, using mortality information alone to construct a health proxy will tend to underestimate the effects of health and overestimate the effects of economic variables on the labor force participation decision, while using self-reported health status can either over- or underestimate the impact of either health or economic variables on such decisions [Bound (1991)].

might imagine trying to estimate θ by non-linear instrumental variables Amemiya (1974)]. However, if g is non-linear not just in parameters, but in variables, this procedure will not consistently estimate θ [Amemiya (1985, 1990), Hsiao (1989)].

For linear models there is a close tie between simultaneous equations and errors in variables models. However, for non-linear models, the analogy breaks down. To see why, imagine that x^* is a linear function of w, $x^* = \pi w + v$, with v orthogonal to w by construction. For the linear model we have:

$$\begin{aligned} y &= x^*\beta + \epsilon \\ &= \pi w\beta + \beta v + \epsilon. \end{aligned} \tag{46}$$

βv is orthogonal to w, so using πw in place of x^* will consistently estimate β. For the nonlinear model we have

$$\begin{aligned} y &= g(x^*;\ \theta) + \epsilon \\ &= g(\pi w;\ \theta) + [g(x^*;\ \theta) - g(\pi w;\ \theta)] + \epsilon. \end{aligned} \tag{47}$$

$[g(x^*;\ \theta) - g(\pi w;\ \theta)]$ will not, in general, be a linear function of v and thus there is no guarantee that it will be orthogonal to $g(\pi w;\ \theta)$[37].

In general, consistent estimation of non-linear errors-in-variables models requires the researcher to know or be able to consistently estimate the conditional distribution of x^* given x, $f(x^* \mid x;\ \delta)$. With f known, the mean of y conditional on x becomes

$$\begin{aligned} E(y \mid x) &= \int g(x^*;\ \theta) f(x^* \mid x;\ \delta)\, \mathrm{d}x^* \\ &= G(x;\ \gamma), \end{aligned} \tag{48}$$

where $\gamma = (\theta, \delta)$. Substituting $G(x;\ \gamma)$ for $g(x^*;\ \theta)$, we obtain a model in terms of observables

$$y = G(x;\ \gamma) + \upsilon, \tag{49}$$

where

$$\upsilon = \epsilon + g(x^*;\ \theta) - G(x;\ \gamma). \tag{50}$$

By construction $E(\upsilon \mid x) = 0$. In principle this model can be estimated by maximum likelihood[38]. Hsiao (1989) proposed computationally simpler minimum distance and two step estimators of the model. Alternatively, one can imagine using multiple

[37] Amemiya (1985) and Hsiao (1989) give more formal versions of this argument.

[38] Simulation techniques can greatly facilitate such estimation [Lavy, Palumbo and Stern (1998), Stinebrickner (1999)].

imputation techniques [Rubin (1987), Little and Rubin (1987), Brownstone (1998)] to first impute estimates of x^* and then use these in a second stage to estimate θ.

The availability of an instrument, w, is not sufficient to allow the researcher to estimate the distribution of x^* conditional on x (or w, for that matter). We have $x = x^* + \mu = \pi w + v + \mu$. The regression of x on w allows us to consistently estimate π, but not the distribution of v. Thus, this first stage regression does not allow us to identify the distribution of x^* conditional on w. Without knowledge of the distribution of x^* conditional on observables, it is not possible to consistently estimates θ. However, the estimator that simply uses $\hat{\pi} w$ as a proxy for x^* often works well [Amemiya (1985), Carroll and Stefanski (1990)] as an approximation[39].

3.4. The contribution of validation data

So far we have been discussing approaches to measurement error that use multiple, possibly error ridden, indicators of the key variables we are interested in, to gauge the reliability of these measures. As we have seen, estimates of the reliability of key measures can be used to gauge the effect of measurement error on our estimates under the assumption that measurement error is, in one way or another, independent of the constructs that enter our models. An alternative is to compare the survey estimate with other, more accurate empirical data. The promise of validation studies is that they give some direct evidence on the nature of the measurement error in survey data, by allowing comparison of survey responses to "true" values if the same variables. Often, the "true" values are obtained from employer or administrative records. Thus, X^* will be referred to as the "record" data.

Consider first the simplest case, where the required validation data is quite modest. Suppose we wish to consistently estimate the effect of a single explanatory variable, x^*, on y^*, but our survey measure for x^* is measured with error. If the error is classical we know $\beta_{yx} = \beta[1 - \sigma_\mu^2/(\sigma_{x^*}^2 + \sigma_\mu^2)]$. Data from a validations study, which includes both the survey response, x, and an accurate measure of x^*, x^r (for example, based on checking reliable administrative records) can give us estimates of σ_μ^2 or $\sigma_\mu^2/\sigma_{x^*}^2$ which can be used to correct the estimate based on the original survey data. Even better, we could not assume the measurement error is classical; as long as it is uncorrelated with y^*, we know that $\beta_{yx} = \beta(1 - \beta_{\mu x})$. The validation data allows us to estimate $\beta_{\mu x}$ directly.

More ambitiously, validation data allows us to identify parameter estimates in the presence of arbitrary patterns of measurement error. Suppose that we have error ridden data for a random (primary) sample of the population. For a distinct random sample of the population we have validation data. We are imagining that this validation data

[39] Amemiya (1985) studies the asymptotic behavior of the nonlinear instrumental variables estimator as σ_μ^2 converges to 0, and finds that with standard regularity conditions, the estimator approaches consistency as σ_μ^2 approaches 0.

contains both the error ridden and error free data. We can then use the validation data to compute the distribution of y^*, X^* given y, X ($f(y^*,X^* \mid y,X)$). This conditional distribution can then be used to impute the distribution of y^* and X^* in the primary data set. What is clearly crucial for such a procedure to be valid is that the distribution of y^*, X^* given y, X be the same in the primary and validation data set [Carroll, Ruppert and Stefanski (1995), refer to this as transportability].

To be somewhat more concrete within the context of the linear model, validation data allow us to calculate empirical analogues to $\beta_{\mu X}$, $\beta_{\nu X}$ and $\beta_{\epsilon X}$, $b_{\mu X}$, $b_{\nu X}$ and $b_{\epsilon X}$. Assume to begin with that one's measure of y in the primary data set is error free and that X is exogenous ($\beta_{\epsilon X} = 0$). Also let $\beta_{X^*,X}$ represent the matrix of regression coefficients from the regression of X^* on X in the validation sample ($\beta_{X^*,X} \equiv \mathrm{I} - \beta_{\mu X}$). A consistent estimate of β can be obtained by first using $\beta_{X^*,X}$ calculated in the validation sample to transform X in the primary sample, $\hat{X} = \beta_{X^*,X}X$, and then regressing y on $\hat{X}$. Note that under these circumstances consistent estimation of β requires validation data on X, but does not require validation data on y. In fact, as the expressions make clear, the validation data on X can come from a separate sample that contains no information on y, as long as both the primary sample and the validation sample are random samples from the same population.

More generally, if $\beta_{\nu X} \neq 0$ and $\beta_{\epsilon X} \neq 0$, then one can obtain consistent estimates of β by transforming y as well as X. Let $\hat{y} = y - [\beta_{\nu X} + \beta_{\epsilon X}]X$. Then

$$\hat{\beta} = \left[\hat{X}'\hat{X}\right]^{-1} \hat{X}'\hat{y}, \tag{51}$$

consistently estimates β[40].

Lee and Sepanski (1995) generalize Equation (51) to the nonlinear context. They consider the nonlinear regression

$$y^* = g(x^*, \theta) + \epsilon. \tag{52}$$

In the primary data set, the researcher has a random sample of error-ridden versions of y^* and x^*, which, following our general notation, we will refer to as y and x. The researcher also has available a validation data set that contains a random sample of both accurately measured and error ridden versions of y^* and x^*, y_{v}, x_{v}, x^*_{v}, y^*_{v}, where the v subscript is used to indicate the data come from the validation data.

Consider first the case where either y^* is accurately measured ($y = y^*$) or where measurement error in y is classical and so can be absorbed in the error term and where the measurement error in x^* is nondifferential. Lee and Sepanski (1995) propose an estimator of θ that minimizes

$$\hat{\theta} \equiv \min_{\theta} \left[y - x(x_{\mathrm{v}}' x_{\mathrm{v}})^{-1} x_{\mathrm{v}}' g(x^*_{\mathrm{v}};\, \theta)\right]^2. \tag{53}$$

They show that under standard regularity assumptions, $\hat{\theta}$ consistently estimates θ and derive its asymptotic distribution. In the context where y^* suffers from non-

[40] These ideas are developed formally and generalized to the non-linear setting in Lee and Sepanski (1995).

classical measurement error or where the measurement error in x^* is differential, Equation (53) can be modified to consistently estimate θ. Define $w = [y : x]$ and $\hat{y} = y - w'(w'_{\mathrm{v}}w_{\mathrm{v}})^{-1}w'_{\mathrm{v}}(y_{\mathrm{v}} - y^*_{\mathrm{v}})$. Then

$$\hat{\theta}' \equiv \min_{\theta} \left[\hat{y} - x(x'_{\mathrm{v}}x_{\mathrm{v}})^{-1} x'_{\mathrm{v}} g(x^*_{\mathrm{v}};\, \theta)\right]^2, \tag{53$'$}$$

will consistently estimate θ.

Measurement error in key variables can be thought of as a special case of missing data – in a literal sense the researcher is missing valid data on the variables measured with error. Much of the voluminous literature on handling missing data has focused on the case where data are missing for a subset of the data. Within the context of measurement error this is akin to having validation data available. Thus, the techniques that have been developed to deal with missing data [Little (1992)] could be applied to estimating models with error ridden data as well[41].

In the general context, the impact of measurement error on parameter estimates is model dependent. As we have seen, within the context of the linear model, the impact will depend on the association between the measurement error in the key variables and all the other variables included in a model. More generally, one needs to be able to estimate $f(y^*, X^* \mid y, X)$, where y^* and X^* include all the variables of interest. Thus, the value of validation studies is enhanced if they include not just data on the key variables being validated but also on other variables that researchers would typically use in conjunction with these variables.

Validation studies report information regarding the magnitude of the measurement error involved in survey measures – typically the mean and some measure of the dispersion in the measure. Correlations between the survey and validation study measures of the same variable will also often be reported and can be thought of as measures of the validity of the survey measures (the validity of a measure is the correlation between the measure and the actual underlying construct that the measure is intended to be a measure of). While information on the marginal distributions of the error is sufficient to allow researchers to use such studies to estimate the impact of measurement error on parameter estimates if measurement error is classical, our discussion should make clear that one of the real values of a validation study is to allow us to relax such assumptions. Studies sometimes report not only summary statistics but also sample regressions. However, even these regressions will provide information regarding the impact of measurement error on estimates only for models similar to the ones reported on in the validation study report. Perhaps such tabulations should be seen as illustrative. While, in general, it will not make sense or be possible to report $f(y^*, X^* \mid y, X)$, it will often be possible to make the validation study data available to

[41] See Carroll, Ruppert and Stefanski (1995) for a discussion of the link between missing data and measurement error models. See Brownstone and Valletta (1996) for the implementation of these ideas within the context of an economic example.

researchers, thus allowing individual researchers to study the impact of measurement error on whatever kind of model they are interested in estimating. Indeed some of the most interesting results in the literature using validation studies have been done by individuals who were not originally involved in collecting the validation data but who use such data to examine the impact of measurement error on parameter estimates within the context of a specific research question.

While validation data has considerable promise, it is important to bear in mind the limitations of such data as well. We have in mind two distinct issues. First validation data presumably has higher validity than survey measures – indeed the very value of such data depends on this presumption – however this does not mean that it is completely without error. Even administrative data or payroll records will include errors. Equally important, validation data may not tap exactly the same construct as does the survey measure and some of the discrepancies between the survey and validation data measures may involve discrepancies between the constructs the two capture. Neither the survey measure nor the validation study measure may adequately capture the construct we are interested in.

Second, validation study data collected in one context, may not generalize to another[42]. In some contexts the issues are obvious. Thus, for example, data collected from a single firm may not be that informative about the nature of measurement error in nationally representative data both because of idiosyncracies regarding the firm and because the data misses any between-firm variation. In other cases, issues are more subtle. Existing methodological work (see Section 5) suggests that for many items the extent of measurement error will be context dependent. For example, the extent of measurement error in reported earnings and employment status appears to depend on the business cycle (see Section 6.1).

4. Approaches to the assessment of measurement error

In order to use the procedures outlined in Section 3, one needs either data that include multiple indicators of variables measured with error or validation data that include both accurate and error ridden versions of the analysis variables. As the above discussion should make clear, the use of multiple measures to correct for biases introduced by measurement error requires the use of strong assumptions about the nature of the

[42] Carroll, Ruppert and Stefanski (1995) emphasize the value of having validation data collected on a random sub-samples of the primary data ("We cannot express too forcefully that if it is possible to construct an internal validation data set, one should strive to do so. External validation can be used … but one is always making an assumption when transporting such models to the primary data."). Validation data collected as a subsample of the primary data is practically nonexistent in the data typically used by economists, but such data has sometimes been collected in other contexts (see the examples discussed by Carroll, Ruppert and Stefanski).

measurement error involved. What is nice about validation data is that it allows the researcher to relax such assumptions.

Multiple indicator or validation data is sometimes collected as part of the primary data collection effort. However, more commonly, such data comes from external independent studies. It should be clear that internal multiple indicator or validation data is to be preferred over external data. With the use of external data one is always making an assumption about the transportability of models from the external to the primary data.

Most of the research involving validation data incorporates one of two designs: (1) obtaining external data for the individuals included in the survey, or (2) comparing external population-based parameters or estimates with those derived from the survey. We examine empirical studies that encompass four separate approaches to the assessment of the quality of household reported economic phenomena:

(i) Validation studies which involve micro-level comparisons of household-reported data with external measures of the phenomena, such as employer's records or administrative records;

(ii) Micro-level comparisons of response variance which involve the comparison of individual survey respondents' reports at time t with reports obtained at time $t+x$, under the same essential survey conditions;

(iii) Micro-level comparisons of response differences involving the comparison of the individual survey respondents' reports at time t with reports obtained at time $t \pm x$, involving either administrative records (e.g., comparison to tax returns) or the collection of survey data under different (and supposedly preferred) survey conditions; and

(iv) Macro-level comparisons of estimates based on survey reports with aggregate estimates generated under different (and supposedly preferred) survey conditions or from aggregate administrative records.

Each of these approaches to the assessment of data quality suffers from potential limitations; these limitations are outlined in the discussion that follows.

Validation studies which permit micro-level comparisons can be classified as one of three types of studies: (1) a reverse record check, in which elements are sampled from the administrative (or validation) records and then interviewed; (2) prospective record checks in which elements are interviewed and then administrative records are checked to confirm the reported behaviors; or (3) complete record check studies, in which all elements in the population have a probability of selection into the sample and administrative records or other validation information are obtained for all sampled elements, regardless of whether the behavior of interest has been reported or not. If the measure of interest is a discrete event (e.g., hospitalization, industrial accident related to a particular job), reverse record check studies are quite adequate in measuring underreporting, but are often insensitive to overreports, since the administrative records may not include the complete universe of events. Prospective record checks attempt to verify affirmative survey responses; thus these designs are better for assessing overreporting of discrete events but less adequate than reverse record check studies for assessing underreporting of events since it may be difficult to obtain validation data from all potential sources. A complete or full record check, provided all of the relevant records can be located, provides the best means for assessing both underreporting as well as overreporting. However, such studies are rare, requiring a closed universe from which one can obtain the validation information and be confident that the records include an accounting for the entire universe of behaviors.

Regardless of the design of the validation study, most empirical investigations incorporating validation data attribute differences between the respondent report and the validation data to the respondent and thus may overstate the level of response error. There are two separate issues here. First, various factors may contribute to measurement error, including the interviewer, the wording of a particular question, the context of the questionnaire, as well as the essential survey conditions such as the mode of data collection; however, differences between survey reports and administrative records are often discussed in terms of response error. Recognizing the alternative sources of errors is a first step in modeling them properly. Second, as noted above, differences between respondent reports and the validation data may reflect deficiencies in the latter. Most record check studies fail to assess or even discuss the level of potential error in the records or the error introduced via the matching of survey and record reports. Comparisons of survey reports with self-reported administrative records (e.g., tax records) may show discrepancies because of errors in the administrative records. Finally, it is rare to see a thorough discussion of the impact of definitional differences between the two sources of information on the level of apparent error.

In contrast to most validation studies, micro-level comparisons of survey reports for discrete events occurring before time t obtained at two points in time under the same essential survey conditions focus on simple response variance over time. However, the accuracy of the data at either time t or time $t+x$ can not be assessed. Empirical investigations of this type usually attribute differences in the two estimates to error in the reports obtained at time $t+x$ (under the assumption that the quality of retrieval declines over time).

Micro-level comparisons which entail survey estimates produced as a result of different survey designs similarly tend to attribute differences in the estimates to response error for the estimates produced under the less optimal design. Hence, the later comparison requires a priori knowledge of the design most likely to produce the most accurate data.

Macro-level comparisons are fraught with several potential confounding factors, including differences in the population used to generate the estimates, definitional differences, and differences in the reference period of interest. Benchmark data are themselves potentially subject to various sources of errors and omissions, complicating the interpretation of differences between the two sources. Finally, whereas micro-level validation can compare survey responses to external data, comparisons of survey data with aggregate benchmark data requires some assumptions about non-response, either reweighting the available responses or imputing values for non-respondents. Perfectly accurate survey responses can appear to diverge from benchmark totals if the handling of non-respondents is in error; incorrect survey responses could even add to correct control totals if response error and errors in nonresponse corrections are offsetting.

5. Measurement error and memory: findings from household-based surveys

The assessment of measurement error across various substantive disciplines has provided a rich empirical foundation for understanding under what circumstances survey responses are most likely to be subject to measurement error. The theoretical framework for most of these investigations draws from the disciplines of cognitive and social psychology. Although these investigations have provided insight into the factors associated with measurement error, there are few fundamental principles which inform either designers of data collection efforts or analysts of survey data as to the circumstances, either individual or design-based, under which measurement error is most likely to be significant or not. Those tenets which appear to be robust across substantive areas are outlined below.

5.1. Cognitive processes

Tourangeau (1984) as well as others [see Sudman, Bradburn and Schwarz (1996) for a review] have categorized the survey question and answer process as a four-step process involving comprehension of the question, retrieval of information from memory, assessment of the correspondence between the retrieved information and the requested information, and communication. In addition, the encoding of information, a process outside the control of the survey interview, determines a priori whether the information of interest is available for the respondent to retrieve from long-term memory.

Much of the measurement error literature has focused on the retrieval stage of the question answering process, classifying the lack of reporting of an event as retrieval failure on the part of the respondent, comparing the characteristics of events which are reported to those which are not reported. One of the general tenets from this literature concerns the length of the recall period; the greater the length of the recall period, the greater the expected bias due to respondent retrieval and reporting error. This relationship has been supported by empirical data investigating the reporting of consumer expenditures and earnings [Neter and Waksberg (1964)]; the reporting of hospitalizations, visits to physicians, and health conditions [e.g., National Center for Health Statistics (1961, 1967), Cannell, Fisher and Bakker (1965), Woolsey (1953)]; reports of motor vehicle accidents [Cash and Moss (1972)], crime [Murphy and Cowan (1976)]; and recreation [Gems, Ghosh and Hitlin (1982)]. However, even within these studies the findings with respect to the impact of the length of recall period on the quality of survey estimates are not consistent. For example, Dodge (1970) found that length of recall was significant in the reporting of robberies but had no effect on the reporting of various other crimes, such as assaults, burglaries, and larcenies. Contrary to theoretically justified expectations, the literature also offers several examples in which the length of the recall period had no effect on the magnitude of response errors [see for example, Mathiowetz and Duncan (1988), Schaeffer (1994)]. These more recent investigations point to the importance of the complexity of the behavioral experience over time, as opposed to simply the passage of time, as the factor most indicative of measurement error.

Another tenet rising from the collaborative efforts of cognitive psychologists and survey methodologists concerns the relationship between true behavioral experience and retrieval strategies undertaken by a respondent. Recent investigations suggest that the retrieval strategy undertaken by the respondent to provide a "count" of a behavior is a function of the true behavioral frequency. Research by Blair and Burton (1987) and Burton and Blair (1991) indicate that respondents choose to count events or items (episodic enumeration) if the frequency of the event/item is low and they rely on estimation for more frequently occurring events. The point at which respondents switch from episodic counting to estimation varies by both the characteristics of the respondent as well as characteristics of the event. As Sudman et al. (1996, p. 201) note, "no studies have attempted to relate individual characteristics such as intelligence, education, or preference for cognitive complexity to the choice of counting or estimation, controlling for the number of events". Work by Menon (1994) suggests that it is not simply the true behavioral frequency that determines retrieval strategies, but also the degree of regularity and similarity among events. According to her hypotheses, those events which are both regular and similar (brushing teeth) require the least amount of cognitive effort to report, with respondents relying on retrieval of a rate to produce a response. Those events which occurred irregularly and which were dissimilar require more cognitive effort on the part of the respondent.

The impact of different retrieval strategies with respect to the magnitude and direction of measurement error is not well understood; the limited evidence suggests that errors of estimation are often unbiased, although the variance about an estimate (e.g., mean value for the population) may be large. Episodic enumeration, however, appears to lead to biased estimates of the event or item of interest, with a tendency to be biased upward for short recall periods and downward for long recall periods. In part, the direction of the estimation error related to episodic enumeration is a function of the misdating of the dates of retrieved episodes of behavior, a phenomenon referred to in the literature as telescoping [e.g., Sudman et al. (1996)]. The evidence for telescoping comes from studies which have examined respondent's accuracy in reporting dates of specific events. Forward telescoping refers to the phenomena in which respondents report the occurrence of an event as more recent than is true; backward telescoping refers to misdating in the opposite direction, that is, reporting the event as occurring earlier in time than is true. The direction of the misdating appears to be a function of the length of the reference period. Forward telescoping is most evident when the reference period is short (one or two weeks), whereas backward telescoping is more common for longer (one year or more) reference periods[43].

The misdating of episodic information in panel data collection efforts has given rise to a particular type of response error referred to as the "seam effect" [Hill (1987)]. Seam effects refer to the phenomena of a disproportionate number of changes in respondent status (e.g., employment status) change at the "seam" between the end of the reference period for wave x of a study and the start of the reference period for wave $x + 1$ of a study. For example, a respondent will report being employed at the time of the wave x interview; at wave $x + 1$, the respondent reports being unemployed for the entire reference period. Hence his or her change in employment status occurred at the seam of the reference periods. Although the seam effect may arise as a function of the misdating of the start or end of a particular status, some have speculated that the effect is a result of respondents minimizing the level of effort associated with the respondent task by projecting the current status back to the beginning of the reference period of interest.

Finally, a third tenet springing from this same literature concerns the salience or importance of the behavior to be retrieved. Salience is hypothesized to affect the strength of the memory trace and subsequently the effort involved in retrieving the information from long-term memory. The stronger the trace, the lower the effort needed to locate and retrieve the information. In a study on the reporting of hospitalizations, Cannell, Fisher and Bakker (1965) found that hospitalizations of longer duration were subject to lower levels of errors of omission than hospitalizations of one or two days in length; Waksberg and Valliant (1978) report a similar pattern with respect to injuries. Although salient information may be subject to lower levels of errors of omission, other research has indicated that salience may lead to overestimation on the part of the respondent [e.g., Chase and Harada (1984)]. As is evident from the literature, overestimation or overreporting on the part of the respondent can result from either forward telescoping of events, that is, the misdating the event of interest counting events which occurred prior to the start of the reference period, or from misestimation, in part, due to the salience of the event of interest. Unfortunately, empirical investigations of response error in which overreporting is evident have not addressed the relative importance of forward telescoping and salience as the source of the response error.

5.2. Social desirability

In addition to asking respondents to perform the difficult task of retrieving complex information from long-term memory, survey instruments often ask questions about

[43] The work on telescoping has focused on the effect of telescoping on the time of individual events. However, it seems likely that when respondents are asked to retrospectively recall the timing of various events in their past, errors in the reported timing of various events are correlated, creating something of a spurious coincidence of events. This is a potentially serious issue for event history analysis.

socially and personally sensitive topics. It is widely believed and well documented that such questions elicit patterns of underreporting (for socially undesirable behavior and attitudes) as well as overreporting (for socially desirable behaviors and attitudes). The determination of social desirability is a dynamic process, a function of the question topic, the immediate social context, and the broader social environment at the time the question is asked. Some topics are deemed, by social consensus, to be too sensitive to discuss in “polite” society. In the 1990s this is a much shorter list than was true in the 1950s, but most would agree that topics such as sexual practices, impotence, and bodily functions fall within this classification. Some hypothesize that questions concerning income also fall within this category [e.g., Tourangeau, Rips and Rasinski (2000)]. Other questions may concern topics which have strong positive or negative normative responses (e.g., voting, the use of pugnacious terms with respect to racial or ethnic groups) or for which there may be criminal retribution (e.g., use of illicit drugs, child abuse).

The sensitivity of the behavior or attitude of interest may affect both the encoding of the information as well as the retrieval and reporting of the material; little of the survey methodological research has addressed the point at which the distortion or measurement error occurs with respect to the reporting of sensitive material. The encoding of emotionally charged behaviors is hypothesized to include an encoding of the emotion associated with the event. The presence of the emotion may affect further retrieval of that information. Cognitive dissonance may lead the respondent to “undo” the details of the event, distorting the event in subsequent rehearsals, thereby encoding the distorted information with the behavior [Loftus (1975)]. Even if the respondent is able to retrieve accurate information concerning the behavior of interest, he or she may choose to edit this information at the response formation stage as a means to reduce the costs, ranging from embarrassment to potential negative consequences beyond the interview situation, associated with revealing the information.

5.3. Essential survey conditions

The measurement process and the quality of survey data can also be affected by design features such as the mode of data collection (e.g., face-to-face, telephone, self-administered), the method of data collection (e.g., paper and pencil, computer assisted interviewing), the nature of the respondent (self vs. proxy response), characteristics of the interviewer (e.g., gender, race, voice quality), cross section vs. longitudinal design, the frequency and time interval between interviews for longitudinal data collection, as well as the data collection organization and survey sponsor. Groves (1989) provides a thorough review of empirical literature related to these various sources of error. While there is evidence that at times, each of these factors may affect the quality of the data, the empirical literature is inconsistent as to the direction and magnitude of the error attributable to each of these design features.

5.4. Applicability of findings to the measurement of economic phenomena

One of the problems in drawing inferences from other substantive fields to that of economic phenomena is the difference in the nature of the measures of interest. As noted earlier, much of the assessment of the quality of household-based survey reports concerns the reporting of discrete behaviors; many of the economic measures that are the subject of survey inquiries are not necessarily discrete behaviors or even phenomena that can be linked to a discrete memory. Some of the phenomena of interest could be considered trait phenomena. Consider the reporting of occupation. We speculate that the cognitive process by which one formulates a response to a query concerning current occupation is different from the process related to reporting number of doctor visits during the past year.

For other economic phenomena, it is likely that individual differences in the approach to formulating a response impact the magnitude and direction of error associated with the measurement process. Consider the reporting of current earnings related to employment. For some respondents, the request to report current earnings requires little cognitive effort – it may almost be an automatic response. For these individuals, wages may be considered a characteristic of their self identity, a trait related to how they define themselves. For other individuals, the request for information concerning current wages may require the retrieval of information from a discrete episode (the last paycheck), a recent rehearsal of the information (the reporting of wages in an application for a credit card), or the construction of an estimate at the time of the query based on the retrieval of information relevant to the request.

Given both the theoretical and empirical research conducted within multiple branches of psychology and survey methodology, what would we anticipate are the patterns of measurement error for various economic measures? The response to that question is a function of how the respondent's task is formulated and the very nature of the phenomena of interest. For example, asking a respondent to provide an estimate of the number of weeks of unemployment during the past year is quite different from the task of asking the respondent to report the starting and stopping dates of each unemployment spell for the past year. For individuals who are in a steady-state (constant employment or unemployment), neither task could be considered a difficult cognitive process. For these individuals, unemployment is not a discrete event but rather may become encoded in memory as a trait which defines the respondent. However, for the individual with sporadic spells of unemployment throughout the year, the response formulation process would most likely differ for the two questions. While the response formulation process for the former task permits an estimation strategy on the part of the respondent, the latter requires the retrieval of discrete periods of unemployment. For the reporting of these discrete events, we would hypothesize that patterns of response error evident in the reporting of episodic behavior across other substantive fields would be observed. Similar patterns of differences may be observed as a function of requesting the respondent to report current earnings as compared to directing them to think about their last paycheck and report the gross earnings. With respect to social desirability, we would anticipate patterns similar to those evident in other types of behavior, overreporting of socially desirable behaviors and underreporting of socially undesirable behaviors.

6. Evidence on measurement error in survey reports of labor-related phenomena

6.1. Earnings

Empirical evaluations of household-reported earnings information include the assessment of annual earnings, usual earnings (with respect to a specific pay period), most recent earnings, and hourly wage rates. Validation data are generally based on employers' or administrative records. Gradually, the focus of such studies has shifted. Early studies tended to focus on whether the mean error was near zero, and so whether the survey reports were unbiased. More recent studies focus on the variance of the error relative to true variation and, more generally, on the bias caused by errors when survey measures of individual earnings are used in linear models. As a result, it is hard to report results from the various studies we review in a consistent fashion. Ideally, we would like to report information on the distribution of errors (e.g., the mean and variance of errors) together with some measure of the potential biases introduced into simple models by the error. Our preferred measure of this potential bias is the slope coefficient from the regression of the record values on the survey values of the same variable. As we have seen, under the assumption that the record values are valid, one minus this coefficient gives a measure of the proportional downward bias introduced by the measurement error for simple bivariate linear regression models that use the variable in question as the explanatory variable[44]. The range of summary measures in Table 1 reflects the considerable variation in what can be computed from studies from different disciplines that are motivated by different questions.

Overall, the findings suggest that annual earnings are reported with less error than hourly wage rates or weekly earnings. Mean estimates of annual earnings appear to be subject to relatively small levels of response error, whereas absolute differences indicate significant over- and underreporting at the individual level. We also find consistent evidence that errors are mean-reverting, but less consistent evidence that errors are correlated with standard human capital and demographic variables.

6.1.1. Annual earnings

Nine of the studies reported in Table 1, representing six different data collection efforts[45], examine the quality of reports of annual earnings. For each of these studies,

[44] The measure will be valid if the employer's or administrative records are valid and error free *or* if the errors in such records are completely random.

[45] The Panel Study of Income Dynamics (PSID) Validation study is represented three times; see Duncan and Hill (1985), Rodgers, Brown and Duncan (1993), and Bound, Brown, Duncan and Rodgers (1994); the CPS-SSA matched study is reported in Bound and Krueger (1991) and Bollinger (1998).

Table 1
Assessment of measurement error: earnings

Reference	Variables of interest	Validation source	Findings
Keating, Paterson and Stone (1950)	Weekly wages on jobs held in previous year (survey of currently unemployed)	Employers' records	*r*(interview, record) = .90 (men) and .93 (women)
Miller and Paley (1958)	Annual earnings (decennial census post-enumeration survey)	IRS tax forms	Receipt of earnings/wages: underreported at 2% to 6%; Comparison of median income indicates small (1%) net bias; underreporting for families (3%), overreporting for unrelated individuals (4%)
Hardin and Hershey (1960)[1]	Weekly earnings (salaried workers)	Employers' records	*r*(interview, record) = .98 for men and .99 for women; Those who had recently received a raise were more likely to underreport their earnings

Borus (1966)[1]	Weekly earnings	Employers' records	Mean (household report): $67.37;
			Mean (employer report): $63.98;
			Mean (simple difference): $3.39;
			r(household, employer) = .95;
			Difference higher for males and those with higher reported earnings and hours; and for older workers and those with more education

Borus (1970)[2]	Annual earnings (comparison of two methods: 2 broad questions concerning earnings and summation of work histories)	Employers' reports of wages to Indiana Employment Security Division	Mean annual earnings: $2500;
			Work history reports: mean error = $46.67, s.d. = $623.49;
			Broad questions: mean error = $38.57, s.d. = $767.14;
			Over 15% of responses misreported $1000 or more;
			Work history approach resulted in smaller response errors among those with some college education and for persons with average or above average earnings while the broad questions resulted in more accurate data among poor persons with no college education

Dreher (1977)[1]	Monthly salary (nine $500 intervals)	Employers' records	*r*(interview, record) = .91

continued on next page

Table 1, *continued*

Reference	Variables of interest	Validation source	Findings

Carstensen and Woltman (1979)	Rate of pay, usual weekly earnings (CPS special supplement)	Employers' report	Those reporting pay per hour:
			Mean (household): $4.21;
			Mean (employer): $4.44;
			Mean (difference): –$.23 (s.e. = $.02);
			Those reporting pay per week:
			Mean (household): $203;
			Mean (employer): $217;
			Mean (difference): –$14 (s.e. = $2.70);
			Those reporting pay per month:
			Mean (household): $1173;
			Mean (employer): $1068;
			Mean (difference): $104 (s.e. = $14.90);
			Those reporting pay per year:
			Mean (household): $16 868;
			Mean (employer): $16 068;
			Mean (difference): $800 (s.e. = $403);
			When pay reported per hour, both self- and proxy reports have small mean error; when pay reported per week, self-reports have small mean error but proxy reports are 20% below true values

Greenberg and Halsey (1983)	Quarterly earnings (participants in Gary and Seattle–Denver income-maintenance experiments)	Employer reports to state unemployment-insurance agency	Controls in S–D experiment slightly overreported earnings whereas Gary controls significantly underreported earnings (28, 37, and 36% for husbands, wives, and female "heads", respectively); Those eligible for experimental income-maintenance payments tended to underreport earnings (except for husbands in S–D); Earnings difference between experimentals and controls exaggerated by misreporting for all groups, ranging from 2–3% of earnings (husbands at both sites) to 16% (young non-heads in S–D)

continued on next page

Table 1, *continued*

Reference	Variables of interest	Validation source	Findings		
Mellow and Sider (1983)[3]	Wage per hour (CPS)	Employers' records	ln (employer reported wage) – ln (worker reported wage): mean = .048; variance = .167; Regression with employer–worker wage difference as the dependent variable, no significant coefficients; Wage equations based on employer vs. worker reported wages indicated no difference in structure of wage determination		
Duncan and Hill (1985)[1]	Annual earnings, year t and $t-1$ (PSID Validation Study)	Employers' records	1982 earnings :	Annual	Hourly
			Mean (interview) :	$29 917	$16.31
			Mean (record) :	$29 972	$16.97
			Mean (difference) :	–55	–.63
			Mean (absolute difference) :	$2313	$2.68
			Error/record variance ratio :	.154	2.801
			1981 Earnings :		
			Mean (interview) :	$29 579	$14.71
			Mean (record) :	$29 873	$15.39
			Mean (difference) :	–294	–.66
			Mean (absolute difference) :	$2567	$2.13
			Error/record variance ratio :	.301	1.835
			1982–1981 Change :		
			Mean (interview) :	$426	$1.61
			Mean (record) :	$179	$1.57
			Mean (difference) :	247	.03
			Mean (absolute difference) :	$2477	$2.82
			Error/record variance ratio :	.501	2.920

continued on next page

Table 1, *continued*

Reference	Variables of interest	Validation source	Findings		

Bound and Krueger (1991)	Annual earnings, previous calendar year (CPS)	Social Security Administration records	Men: annual earnings (household report): mean = $15 586;
			ln interview earnings – ln record earnings: mean = .004, variance = .114;
			Women: annual earnings (household report): mean = $7906;
			ln interview earnings – ln record earnings: mean = –.017, variance = .051;
			(above based on sample with record earnings below SS maximum)		

			1977 ln (earnings) :	Men	Women
			variance (interview)	.437	.666
			variance (record)	.529	.625
			variance (difference)	.116	.051
			r(interview, record)	.884	.961
			r(error, record)	–.420	–.028
			b(record on interview)	.974	.962
			1977–1976 change in ln (earnings) :		
			variance (interview)	.186	.437
			variance (record)	.223	.394
			variance (error)	.121	.089
			r(interview, record)	.707	.894
			r(error, record)	–.481	–.123
			b(record on interview)	.775	.848
			Mismeasurement of earnings leads to little bias when CPS earnings on left-hand side of regression (errors weakly related to regressors); positive autocorrelation between errors in CPS reported earnings; coefficient of .40 for men and .10 for women		

continued on next page

Table 1, *continued*

Reference	Variables of interest	Validation source	Findings		
Coder (1992, Tables A, B)	Sum of husband's and wife's annual wage and salary income (SIPP)	Matched tax return information (from joint returns)	For sample with unimputed SIPP earnings:		
				Earnings	ln (earnings)
			mean (interview)	$40 030	
			mean (record)	$42 060	
			variance (interview)	787×10^6	1.290
			variance (record)	1446×10^6	.822
			variance (error)	454×10^6	
			r(interview, record)	.834	
			r(error, record)	–.687	
			b(record on interview)	1.130	
			Mean error larger (in absolute value) when SIPP earnings are partially of completely imputed.		

continued on next page

Table 1, *continued*

<table>
<tr><th>Reference</th><th>Variables of interest</th><th>Validation source</th><th>Findings</th></tr>
<tr><td>Rodgers, Brown and Duncan (1993)[1,3]</td><td>Earnings and hourly wage (each measured three ways): annual; most recent pay period; usual (PSID Validation Study)</td><td>Employers' records</td><td>Correlation between interview and record
<table>
<tr><th></th><th>ln Earnings</th><th>ln Hourly wage</th></tr>
<tr><td>Annual</td><td>.784</td><td>.651</td></tr>
<tr><td>Most recent</td><td>.675</td><td>.437</td></tr>
<tr><td>Usual</td><td>.456</td><td>.258</td></tr>
</table>
Correlation between error and record
<table>
<tr><th></th><th>ln Earnings</th><th>ln Hourly wage</th></tr>
<tr><td>Annual</td><td>–.216</td><td>.066</td></tr>
<tr><td>Most recent</td><td>–.301</td><td>–.150</td></tr>
<tr><td>Usual</td><td>–.436</td><td>–.191</td></tr>
</table>
Wage rates calculated from reported earnings and hours; variance of the errors can be decomposed into three parts: variance in errors in reported earnings, variance in errors in reported hours, and minus the covariance of those two errors. For annual wage rates, contribution due to error in annual earnings and annual hours are about equal (.93 and .80); errors are positively correlated (r = .43); covariance is negative (–.74); for wage rate based on most recent pay period, errors in reported earnings are about twice as important as errors in reported hours (1.36 and .62); covariance again is negative (–.98). Based on usual pay the estimates are 1.26, .32, and –.58;
Mean error significantly different from zero (albeit small); significantly related to true values (negative), impact the magnitude of regression coefficients when wages are on the left hand side of the equation; and are correlated (weak, positive) across time</td></tr>
</table>

continued on next page

Table 1, *continued*

Reference	Variables of interest	Validation source	Findings		
Bound, Brown, Duncan and Rodgers (1994)[1,3]	Annual earnings and wage per hour (PSID Validation Study); wage per hour for hourly workers only	Employers' records	Mean differences between record and interview values of ln (earnings) small and statistically insignificant for both annual earnings and hourly wage		
			1986	ln Earnings	ln Wage/hour
			variance (interview)	.0488	.0204
			variance (record)	.0416	.0085
			variance (difference)	.0108	.0121
			r(interview, record)	.8862	.6350
			r(error, record)	–.0785	–.0109
			b(record on interview)	.8180	.4085
			1986–1982 change in :		
			variance (interview)	.0365	.0433
			variance (record)	.0357	.0112
			variance (difference)	.0164	.0376
			r(interview, record)	.7738	.3786
			r(error, record)	–.3219	–.1404
			b(record on interview)	.7657	.1930
Branden and Pergamit (1994)	Starting wages (NLSY-79)	Respondent's reports in year t compared to year $t+1$	Mean absolute difference in ln(starting wage) = .10; 42% of respondents report the same starting wage at the two points in time. Consistency related to the time unit used for reporting, with the highest rate of consistency among those reporting starting wage as an hourly or daily rate (47% and 52%, respectively); lowest among those reporting a biweekly wage rate (13% consistent)		

continued on next page

Table 1, *continued*

Reference	Variables of interest	Validation source	Findings
Barron, Berger and Black (1997, Table 5.1)	Starting wages (Upjohn Institute Survey)	Employers' records	Mean (interview) = \$8.84; Mean (record) = \$9.95; difference in means not significant (t = 1.31); r(interview, record) = .974
Bollinger (1998)	Annual earnings, previous calendar year (CPS)	Social Security Administration records	Measurement error more severe (larger mean error for men, larger error variance for women) in single cross-section than in two-year panel; Negative relationship between error and true value (for men) driven by small number of cases with low record earnings; Median error unrelated to true earnings at all levels of earnings
Angrist and Krueger (1999)[3]	Hourly wage (CPS)	Employers' records	ln (employer reported wage) – ln (employee reported wage): mean .017 variance (interview) .355 variance (record) .430 variance (difference) .238 r(interview, record) .650 r(error, record) –.489 b(record on interview) .770 Recoding lowest (highest) one percent of employee-reported wages to the 1st (99th) percentile value increased b(record on interview) to .88

[1] Sample limited to a single employer.

[2] Small sample (n = 300) in single geographic area; assessment of the accuracy of reports of annual earnings based on 173 persons for whom household reports could be linked to state employment records.

continued on next page

Table 1, *continued*

Reference	Variables of interest	Validation source	Findings

[3] Respondent not asked to report hourly wage rate directly. Wages (Mellow and Sider; Angrist and Krueger) or hourly earnings (Rodgers, Brown and Duncan, and Bound, Brown, Duncan and Rodgers) calculated from earnings divided by hours worked; error in reported hours therefore contributes to error in hourly wage rate.

comparisons are made between the survey reports and either administrative data (IRS tax forms, Social Security Administration records) or employers' records.

Miller and Paley (1958) compared 1950 Census reports and IRS data for a sample of Census respondents[46]. Limiting attention to families for which each member over age 14 could be matched to an income tax report (including spouses on joint filings)[47], they found that median earnings were $3412 in the Census reports and $3570 in the IRS data. Moreover, the two distributions appear quite similar (see Table 10 in the original paper). While Miller and Paley do not ask whether the errors in the Census reports are mean reverting, the similarity of the two distributions suggests they must be.

By focusing on IRS records for validation, Miller and Paley excluded those with earnings low enough that they do not file an income-tax report. In contrast, Borus (1970) focused on survey responses of residents in low-income Census tracts in Fort Wayne, Indiana. He experimented with two methods for collecting annual earnings from respondents, a set of two relatively broad questions concerning earnings and a detailed set of questions concerning work histories. The responses from both sets of questions were compared to data obtained from the Indiana Employment Security Division for employment earnings covered by the Indiana Unemployment Insurance Act (e.g., excludes agricultural employees, self-employed, and those working for relatives). The mean annual earnings among the respondents was $2500; although the mean error for the two question types was relatively small, $47 and $39 for the work history and broad questions, respectively, the standard deviation of the mean error was large ($623 and $767). Over 10% of the respondents misreported annual earnings by $1000. While these individual-level errors seem large relative to the mean values, they are similar in magnitude to more recent estimates based on nationally representative samples [e.g., Bound and Krueger (1991)].

In contrast to one of Borus's conclusions, Smith (1997) finds that, among low-income individuals eligible to participate in federal training, earnings data based on adding up earnings on individual jobs leads to significantly higher values than data based on direct questions about annual earnings. In Smith's data, this difference is due to higher values for hours worked and for irregular earnings (overtime, tips, and commissions). Comparisons with administrative data for the same individuals lead Smith to conclude that the estimates based on adding up earnings across jobs leads to overreporting, rather than more complete reporting.

Carstensen and Woltman (1979) compared reports of annual earnings obtained in a special supplement to the January (1977) Current Population Survey (CPS) with

[46] Of 7091 families, only 3903 were completely matched. One important reason for non-matches is income low enough that no Federal tax would be owed; except for this difference, Miller and Paley (1958) find the matched sample representative of the larger Census sample.

[47] Note that for families with more than one earner, we are really comparing family rather than individual earnings.

employers' reports. Respondents in rotation group 7 (1/8 of the entire CPS sample)[48] were asked to report earnings as well as report his or her employer's complete name and address. While one of the major strengths of the design is the nationally-representative sample of household respondents, the effective response rate of 61% raises questions as representativeness of the sample of matched employer–employee information[49]. The use of a mail questionnaire to obtain information from the employer suggests that comparisons between the employer and employee information must consider measurement (or reporting) error in the validation data as a potential source of the discrepancy. The study includes comparisons of annual earnings, as well as hourly, weekly, and monthly rates of pay and usual hours worked. With respect to annual earnings, the absolute difference in the two earnings sources was $800 (s.e. = $403), or about 5% of the mean annual earnings[50].

The Panel Study of Income Dynamics (PSID) Validation Study consisted of two waves of interviews with respondents sampled from a single large manufacturing firm and the corresponding record information for those respondents[51]. Cooperation by the firm essentially eliminated problems of matching validation data to each respondent and allowed for the resolution of anomalies in the validation data. The questionnaire used at both waves requested that the respondent provide information for the previous two calendar years. At the time of the initial interview (1983), the firm's hourly workforce was fully unionized and virtually all employees, both hourly and salaried, worked full-time. The workforce was considerably older and had more job tenure than was true of national sample of workers, in part due to layoffs and few new hires in the two years prior to the initial interview. These deviations were offset by a sampling procedure that disproportionately sampled younger and salaried workers. Comparisons between the two validation samples and data from the Panel Study of Income Dynamics for the respective years indicates that, with respect to annual and hourly earnings, the validation sample respondents have considerably higher means and lower variance than a national sample.

[48] Note that each rotation group of the CPS sample is a nationally-representative sample.

[49] Of the 6791 eligible persons in the CPS, 5591 (82%) provided complete employer address data. Among the employers for whom address information was provided by the CPS respondent, 4166 (75%) responded to the mail survey which included the same earnings and hours questions asked of the CPS household respondent, resulting in an effective response rate of 61%.

[50] Respondents in the Carstensen and Woltman study could report earnings in the time unit of their choice, that is, annual, weekly, monthly, or hourly. The comparison of annual earnings was limited to those respondents for whom both the respondent and the employer reported the earnings as annual earnings.

[51] The PSID-VS was conducted by telephone with workers at their homes, rather than administered at the workplace. Similar to other household-based studies, the PSID-VS suffered from nonresponse. The initial wave of interviewing was conducted in the summer of 1983 with 418 of the 534 sampled employees (78.3%). A second wave of interviewing was conducted in the summer of 1987. The sample consisted of respondents to the initial wave and a fresh sample of hourly workers; the response rate among the initial wave respondents was 82.4% and 74.7% among the new sample of hourly workers, resulting in an overall sample size of 492 completed interviews.

Using data from the first validation study, Duncan and Hill (1985) compared reports of annual earnings for calendar year 1981 and 1982 with information obtained from the employer's records. For neither year is the mean of the simple difference between the two data sources statistically significant, although the absolute differences for each year indicate significant under- and overreporting. The average absolute difference between the interview and record reports of earnings for 1982 was \$2123, approximately 7% of mean earnings. The report of earnings for 1981 was of lower quality than for 1982; the absolute difference of the two reports of earnings for 1981 was \$2567, or approximately 8.5% of mean earnings. The error-to-true variance ratio showed a larger difference between the two years: for calendar year 1982 annual earnings it was quite small (.154) but significantly larger for 1981 (.301).

While the margin of difference depends on the measure employed, by all indications previous-year's earnings are reported more accurately than those of two years prior to the interview. While this is consistent with greater error for longer recall periods, it may also reflect the fact that 1981 was a year of economic disruption both for the economy and for this firm.

Comparison of measures of change in annual earnings based on the household report and the employer records indicate no difference in means. Error to true variance ratios are higher for changes than for levels (.50 vs. .15–.30), even though mean absolute errors are similar for changes and levels. Errors in reported changes would be higher but for the positive correlation between the errors in the two years, .43. Duncan and Hill (1985) emphasize that these changes are obtained from differencing reports for two calendar years in the same interview, not differencing reports of last year's earnings from two interviews in a longitudinal survey.

Although the findings noted above are often based on small samples drawn from either a single geographic area [Borus (1970)] or a single firm [Duncan and Hill (1985)], the results parallel the findings from nationally representative samples. Bound and Krueger (1991) created a longitudinal linked file based on the 1977 and 1978 March CPS questionnaires and earnings histories from Social Security Administration files[52]. The study is restricted to those respondents classified as heads of households for whom information for March of 1978 was successfully matched to data reported in March of 1977 and the Social Security records. Of the 27 485 persons classified as heads of households in matchable rotation groups (50% of the CPS rotation groups), the three-way link was made for 9137 persons. Other limitations (e.g., private, covered employment and positive (non-imputed) CPS and SSA earnings in both years) further reduced the effective sample to approximately 3500 persons. Bound and Krueger note

[52] As part of a joint project of the Census Bureau and the Social Security Administration, survey responses for persons in the March (1978) CPS Annual Demographic File were linked to their respective earnings information in SSA administrative records to create the CPS-Social Security Records Exact Match File (CPS-SER). To create a longitudinal data set, the CPS-SER was matched to the (1977) March CPS Annual Demographic File, based on the respondent's unique CPS identification number, age, education, sex, and race.

that the matching process tends to eliminate those who misreport their Social Security number or other matching data, and so those who tend to give inaccurate responses to other questions (e.g., earnings) may be under-represented. Another caveat is that the Social Security earnings data refer to earnings taxable under the payroll tax, and nearly half of the males in their sample reach this limit. Consequently, many of the estimates reported below are based on models that correct for this truncation, based on the assumption that ln earnings are normally distributed.

Bound and Krueger (1991) examined error in annual ln earnings reports separately for men and women. Although the error was distributed about a near-zero mean for both men and women, the magnitude of the error was substantial. For men, the error variance exceeded .10 and represented 27.6% of the total variance in CPS earnings; for women the error variance was approximately .05 and represented less than 9% of the total variance in CPS earnings for women. One striking feature of the errors is that while they appear to be unimodal and symmetric, the tails are substantially thicker than one would expect with a normal distribution. Indeed, for those for whom errors were directly observable (those below the Social Security earnings limit), the standard deviation of the errors was three times the interquartile range. In addition the distributions show a large spike near 0. For those below the earnings limit, 12% of men and 14% of women report earnings that exactly match their Social Security records, while more than 40% of each gender report earnings within 2.5%.

Despite these errors, the correlation between interview and record ln earnings is high in Bound and Krueger's data (.88 for men and .96 for women). Errors are negatively related to the record value for men (−.42), and essentially uncorrelated for women (−.03). Because errors for men are mean-reverting and errors for women are small, they find that measurement error should not appreciably bias the coefficient of ln earnings in linear models. The regression of record on interview values gives coefficients very close to 1 (.97 for men and .96 for women).

Because their data include two CPS waves, they can compare interview and record reports of changes in earnings as well. Consistent with the conventional wisdom, differencing increases the error variance (from .1 to .12 for men, and from .05 to .09 for women), and reduces the true variance by about half. Positive correlation in the errors (.4 for men, .1 for women) limits the increase in error variance due to differencing. Consequently, although the ratio of error to total variance is substantial (.65 for men, .2 for women) the regression of record changes on interview changes (.77 and .85) suggest that the bias due to measurement error when the change in ln earnings is an explanatory variable is not overwhelming.

Bollinger (1998) extended the work of Bound and Krueger (1991), examining the measurement error associated with each of the cross-sectional samples encompassing Bound's and Krueger's panel sample, expanding the sample to include women who were not heads of households, and incorporating nonparametric estimation procedures. To a large extent, Bollinger's findings confirm those of Bound and Krueger. In addition, he finds higher measurement error in the cross-section samples as compared to the panel used by Bound and Krueger, suggesting that constructing a panel from CPS lead to the selection of respondents who appear to be better reporters. Bollinger also finds that the negative correlation between measurement error in reports of annual earnings and record earnings appears to be driven by a small proportion of men with low income who grossly overreport their earnings – or whose earnings are largely unrecorded by Social Security. Of additional interest in the work by Bollinger is the finding that although mean response error is negatively related to earnings, median response error is zero across earnings levels, suggesting median wage regression to be more robust to the effects of response error.

Coder's (1992) analysis compares reports by respondents to the Survey of Income and Program Participation and Federal tax returns. The study is limited to SIPP respondents who were married couples as of March 1991, who met the following criteria: (1) valid Social Security numbers were reported for both the husband and the wife; (2) the couple could be matched to a married-joint tax return; and (3) nonzero wage and salary income amount was reported either during the SIPP interview or on the tax return. Of the approximately 9200 husband–wife couples in the SIPP, 62% (or approximately 5700 couples) met the criteria. Coder finds little difference between mean estimates of annual earnings and the respective validation source. He reports a simple correlation between earnings reported in SIPP and IRS data as .83; the mean annual earnings based on SIPP averaged approximately 4% less than the mean based on matched tax records. Coder's data has an unusually large discrepancy between the variance of interview and record data, with the former smaller; this in turn implies a very strong negative correlation between the "error" (SIPP–IRS) and the "true" (IRS) value – so strong that the effects of earnings on other variables would be overstated due to (mean-reverting) errors in earnings. Alternatively, it is possible that errors in the IRS data contribute to these results [Rodgers and Herzog (1987, p. 408)].

Bound, Brown, Duncan and Rodgers (1994) analyze data from both 1983 and 1987 waves of the PSID Validation Study. The correlation between interview reports and company-record data on ln earnings is about .9 (.92 for 1982 earnings, .89 for 1986), but the negative correlation between error and record values is weaker for 1986 (−.08 vs −.30). Consequently, the regression of record on interview value is closer to 1.0 for 1982 than for 1986 (.96 vs .82).

The distribution of errors for the PSID validation study appear to be quite different than that found by Bound and Krueger (1991) using the matched CPS–Social Security Earnings data. Since virtually all the individuals in the PSID validation study are men, it seems natural to compare PSID validation study results to those for men using the CPS–SSE matched data. The two error distributions have similar means and interquartile ranges, but the PSID validation study data shows neither the spike at 0 nor the thick tails shown by the CPS–SSE matched data. As a result of the thick tails in the CPS–SSE data the variance of errors is an order of magnitude larger in the CPS–SSE data than it is in the PSID validation study data! What accounts for the difference in the distribution of errors between the PSID validation study and CPS–SSE data is unclear [see Bound, Brown, Duncan and Rodgers (1994, p. 357) for a further discussion of these issues].

Given that the change in ln earnings computed from the PSID Validation Study covers four years rather than one, the findings for this variable should be seen as complementing rather than replicating Bound and Krueger's. The general patterns are strikingly similar – increased error variance, with the increase somewhat limited by the correlation in the errors over time; negative correlation between the error and the true value of the change (−.32), and regression of true change on interview reports of .77. One interesting difference is that the correlation between the errors is lower in Bound et al.'s (1994) data (.14) than in Bound and Krueger's (1991). To some extent, this might be expected if the factors that produce this error change gradually over time; on the other hand, it may also reflect the difference between economy wide and single firm samples.

Three of these studies – Duncan and Hill (1985), Bound and Krueger (1991) and Bound, Brown, Duncan and Rodgers (1994) – explore the implications of measurement error for earnings models. Duncan and Hill's model relates the natural logarithm of annual earnings to three measures of human capital investment: education, work experience prior to current employer, and tenure with current employer, using both the error ridden self-reported measure of annual earnings and the record-based measure as the left-hand-side variable. A comparison of the ordinary least squares parameter estimates based on the two dependent variables suggests that measurement error in the dependent variable has a sizeable impact on the parameter estimates. For example, estimates of the effects of tenure on earnings based on interview data were 25% lower than the effects based on record earnings data. Although the correlation between error in reports of earnings and error in reports of tenure was small (.05) and insignificant, the correlation between error in reports of earnings and actual tenure was quite strong (−.23) and highly significant, leading to attenuation in the estimated effects of tenure on earnings based on interview information.

Bound and Krueger (1991) also explore the ramifications of an error-ridden left-hand-side variable by regressing error in reports of earnings on a number of human capital and demographic variables, including education, age, race, marital status, region, and SSA. Similar to Duncan and Hill (1985), the model attempts to quantify the extent to which the correlation between measurement error in the dependent variable and right-hand-side variables biases the estimates of the parameters. However, in contrast to Duncan and Hill, Bound and Krueger conclude that mismeasurement of earnings leads to little bias when CPS-reported earnings are on the left-hand-side of the equation.

Bound, Brown, Duncan and Rodgers (1994) estimate separate earnings functions using both interview and record earnings for both waves of the Validation Study. They find some evidence that errors in reporting ln earnings are negatively related to tenure in 1982, and positively related to education in 1986. Overall, though, they find no consistent pattern. Rodgers, Brown and Duncan (1993) note, however, that if annual hours are included as an explanatory variable, its coefficient is severely biased by a number of factors (e.g., correlation between errors in reporting hours and earnings, in addition to problems with the reliability of hours per se, as discussed in Section 6.1.2).

While there is not much evidence that errors in reported earnings are strongly related to standard explanatory variables in earnings functions, two cautions should be noted. First, the tendency for errors in reported earnings to be mean-reverting means that, if there are no other problems, coefficients of all explanatory variables are biased toward zero. This bias is about 20% of the true coefficient in both studies. Second, errors in other variables may be correlated with earnings, but there is very little evidence one way or the other on this score.

The CPS–SSA matched data and the PSID validation data can also be used to shed some light on the impact of measurement error on earnings dynamics. The short nature of the CPS–SSA matched data panel limits its usefulness for this purpose, but the PSID validation study includes a total of six years of data. Using these data Pischke (1995) found that a relatively simple model in which measurement error in earnings stems from the under reporting of transitory earnings fluctuations together with a white noise component did a good job of explaining basic patterns in the PSID validation study data[53].

Pischke's model rationalizes a number of the stylized facts that have emerged from recent earnings validation studies. In particular his model accounts for the finding that despite mean reversion, measurement error in earnings does not seem to significantly bias the coefficients on explanatory variables in earnings regressions – the explanatory variables in such regressions would be expected to explain permanent, but not transitory earnings.

In terms of the estimation of earnings dynamics, Pischke's estimates imply relatively good news. The negative correlation of measurement error with transitory earnings attenuates the role of the white-noise component. Pischke estimates that surveyed earnings tend to exaggerate the actual fluctuation in earnings by between 20 and 45% depending on the year, but do a reasonably good job identifying the relative importance of the permanent component to earnings changes[54].

There are a few things that are important to note about the Pischke study. First, his model implies reporting errors will tend to be more severe at some points in time as against others (i.e., reporting errors will tend to rise in magnitude as the transitory component of earnings rises). Second, as Pischke emphasizes, it is hard to know how to generalize his results to more representative samples. Even were the PSID validation study establishment representative of establishments in the country as a whole, earnings

[53] With 11 free parameters, Pischke fits 28 free covariances quite well. He reports an overall chi-square statistic on the model of 23.8 (p-value: 0.124).

[54] It is certainly possible to doubt the general validity of Pischke's conclusion. His estimates are based on a tightly parameterized model that was estimated on data from a single firm. However, Baker and Solon (1998) have recently estimated earnings dynamic patterns using administrative data that are remarkably similar to patterns other authors [Baker (1997), Haider (2001)] have found using survey data. These estimates would seem to confirm Pischke's finding that measurement error does not have dramatic effects on estimated earnings dynamics.

dynamics in the sample would miss the component that arises when individuals move across firms.

On balance, the validation evidence suggests little bias in estimating mean annual earnings, and this is quite consistent with the fact that survey-based estimates of earnings aggregated up to economy-wide estimates correspond quite closely to earnings as measured in the National Income and Product Accounts[55]. Moreover, despite significant absolute differences between household reports and record reports of earnings as well as significant error to record variance ratios, the correlation between the various sources of data are quite high. Several of the studies indicate coefficients for the regression of household reports on record reports of annual earnings near 1.0, reflecting a negative correlation between error in the household reports and the record value for annual earning. Only one study addressed the deterioration of the quality of reports of annual earnings as a function of time [Duncan and Hill (1985)]; similar to empirical investigations in other fields, their findings provide support for less accurate reporting for longer reference periods. The evidence with respect to the impact of error in household reports of earnings is mixed; Duncan and Hill (1985) report significant attenuation in a model examining the effects of human capital investment, whereas Bound and Krueger (1991) conclude that misreporting of earnings leads to little bias for models incorporating CPS-earnings on the left-hand-side of the equation.

What can account for the significant individual differences between household and record-reported annual earnings? The reporting of annual earnings within the context of a survey is most likely aided by the number of times the respondent has rehearsed the retrieval and reporting process for this information. We contend that the memory for one's annual earnings is reinforced throughout the calendar year, for example, in the preparation of federal and state taxes or the completion of applications for credit cards and loans. To the extent that these requests have motivated the respondent to determine and report an accurate figure, such information should be encoded in the respondent's memory. Indeed, both CPS and PSID time their collection of annual earnings data to coincide with the time when households would have received earnings reports from employers and might have begun preparing their taxes. Subsequent survey requests should therefore be "routine" in contrast to many of the types of questions posed to a survey respondent. Hence we would hypothesize that response error in such situations would result from retrieval of the wrong information (e.g., annual earnings for calendar year 1996 rather than 1997), social desirability issues (e.g., overreports related to presentation of self to the interviewer), or privacy concerns, which may lead to either misreporting or item nonresponse.

However, several cognitive factors may affect the quality of reports of annual earnings. Comprehension may impact the quality of the information; for example,

[55] For example, CPS-based estimates of total wage and salary income were 97% of independent estimates based on NIPA in 1990 [U.S. Census Bureau (1993)]. As noted earlier, such a comparison reflects several factors besides the mean level of error in the individual reports, such as the accuracy of CPS adjustments for non-response.

respondents may misinterpret the request for earnings information as a request for net earnings as opposed to gross earnings. In addition, the wording of most earnings questions does not stress the need for the reporting of exact earnings; hence respondents may interpret the question as one in which they are to provide estimates as opposed to precise reports of earnings. Estimation on the part of the respondents, as noted by Sudman, Bradburn and Schwarz (1996), often leads to reports that are noisy at the individual level but unbiased at the population level. Retrieval of earnings information for any one year may also be subject to interference with respect to stored information concerning earnings in previous years. If the source of the misreporting by respondents was due to social desirability bias, we would anticipate that the direction of the error would be toward overreporting of annual earnings, especially among those with low levels of earnings and possibly, underreporting among those at the highest levels of earnings. Although there is evidence of a negative correlation between response error and the true value overall, there is little evidence to support the existence of social desirability bias with respect to the reporting of annual earnings [e.g., Bollinger (1998)].

6.1.2. Monthly, weekly, and hourly earnings

In contrast to the task of reporting annual earnings, the survey request to report most recent earnings or usual earnings is more likely to be a relatively unique request and one which may involve the attempted retrieval of information that may not have been encoded by the respondent, the retrieval of information that has not been accessed by the respondent before, or the calculation of an estimate “on the spot”. Hence, we would anticipate that requests for earnings in any metric apart from a well-rehearsed metric would lead to significant differences between household reports and validation data. Moreover, the extent of rehearsal is likely to differ by type of worker; for example, those paid a monthly salary are more likely to have accessed information about monthly earnings than are those paid by the hour, while the reverse is likely for earnings per hour.

While annual earnings is the most frequently studied measure of labor market compensation in validation studies, Table 1 makes it clear that significant effort has also been devoted to validating other measures. Roughly speaking, we can divide these studies into two groups: those that study weekly or monthly pay, and those that study pay per hour.

Four of the earliest studies in Table 1 focus on the correlation between weekly or monthly earnings as reported by workers and their employer’s reports. All four (Keating, Paterson and Stone’s (1950) study of jobs held in the past year by unemployed workers in St. Paul; Hardin and Hershey’s (1960) study of salaried workers at an insurance company; Borus’s (1966) study of average weekly earnings of training-program participants; and Dreher’s (1977) study of average salary of workers at an oil company) report correlations of .90 or higher. Mean reports by workers are close to record values, with modest overreporting in some studies and underreporting in others.

Broadly speaking, these results parallel those reported above for annual earnings, except that the issues of mean reversion and accuracy of changes in panel surveys were not addressed[56].

Carstensen and Woltman (1979) compare worker and employer reports, using a supplement to the January (1977) CPS. Their survey instruments allowed both workers and employers to report earnings in whatever time unit they preferred (e.g., annual, monthly, weekly, hourly). As noted earlier, comparisons are limited to those reports for which the respondent and the employer reported earnings using the same metric. Curiously, when earnings are reported by both worker and employer on a weekly basis, workers underreport their earnings by 6%; but when both report on a monthly basis, workers overreport by 10%. When the various reports are converted to a common time unit (usual weekly earnings), they find workers report earning 11.7% less per week than their employers' reports. Unfortunately, they do not report correlations between worker and employer reports.

Studies of hourly wages or earnings per hour are less common, in part because it is difficult to obtain validation data for salaried workers. Typically, their pay is stated in weekly, monthly, or annual terms, and employers often do not have records of the weekly hours of their salaried workers (see Section 6.4).

In their study of wages, Mellow and Sider (1983) utilized the January (1977) CPS data first analyzed by Carstensen and Woltman (1979)[57]. Hourly wages calculated from the CPS reported earnings and hours compared to employers' records indicate a small, but significant, rate of underreporting (ln hourly wage as reported by the worker lower by .048). The variance of the difference between interview and record reports is .148, which is larger than Bound and Krueger's error variances for the logarithm of annual earnings in CPS data (.114 and .051 for men and women).

In a reanalysis of the same data used by Mellow and Sider, Angrist and Krueger (1999) report more details. In their basic sample they find that the variance in the difference between interview and record values of ln hourly earnings to be .24. In comparison they report the variance in the ln of survey earnings to be .36. While the ratio of these two numbers suggests a signal to total variance ratio of one third, Angrist

[56] Keating, Peterson, and Stone show a cross-tabulation of interview vs. record reports which displays at least weak mean reversion for men. However, from their grouped data, in which 70% of the 115 cases are on the diagonal, it is hard to say anything more precise.

[57] In the CPS sample, validation data could be obtained only where the worker provided the name and address of the employer, and the employer provided the relevant data. Mellow and Sider note that validation data could be obtained for only about two thirds of the eligible sample. However, reported CPS earnings of those who refused to provide employer contact, or whose employers refused to provide validation data, were similar to earnings of those who did not refuse. The EOPP was actually two large studies: a survey of approximately 5000 establishments and the other of approximately 30 000 households. Because of the geographic overlap between the two studies, it was possible to link a limited number ($n = 3327$) of worker and employer responses. The representativeness of the resulting sample is unclear, and was not discussed by Mellow and Sider (1983).

and Krueger's tabulations suggest very substantial mean reversion. The regression of record on surveyed ln earnings suggests attenuation of about 25%.

Duncan and Hill's (1985) analysis of PSID Validation Study data investigates the accuracy of earnings per hour values calculated from workers' reports of annual earnings, weeks worked, and average hours per week. Because hours data were available only for hourly workers, their analysis excludes the firm's salaried workers. On average, calculated earnings per hour are relatively accurate (underreported by about 4%). But the error to true variance ratio of 2.8 leads the authors to characterize the extent of measurement error as "enormous" – the unhappy result of annual earnings being less accurately reported for hourly than for salaried workers and substantial error in reports of annual hours (see below).

Bound, Brown, Duncan and Rodgers (1994) report similarly discouraging results for the logarithm of earnings per hour – error to true variance ratios of about 1.5 in both 1982 and 1986, and correlations between interview and record values of .51 and .64. Predictably, matters only get worse for the change in the logarithm of earnings per hour.

The correlations between interview and record values are strikingly lower than those for weekly or monthly earnings in company-based samples noted above. The earlier company-based studies focused on salaried workers, whereas the PSID Validation Study's hourly earnings information is available only for hourly workers. As it happens, these workers are unionized and the number of hours per week is relatively compressed. In a sense, the poor results for hourly pay occur not because the reporting errors are so large (the standard deviations of the errors are .11 and .16 in the two years) but because true variation is so limited (standard deviations of .09 and .13).

Rodgers, Brown and Duncan (1993), using data from the second wave of the PSID validation study, analyze the accuracy of the logarithm of reported earnings and calculated earnings per hour over three time intervals – annual, most recent pay period, and "usual"[58]. Their analysis is restricted to hourly workers, since record data on hours per week were unavailable for salaried workers. Two generalizations are evident from Table 1: the correlation between worker and record reports declines as one moves from annual to pay period to "usual"; and for any given time interval, earnings per hour are less accurately reported than earnings.

Since wage rates were calculated from reported hours and earnings the variance in the error associated with the wage rate can be decomposed into three parts: the variance of the error in reported earnings, the variance of the error in reported hours, and the covariance of those two reports. While the details vary with the time interval, in general all three of these components are important[59].

[58] Operationally, they define "usual" as the average over the preceding six two-week pay periods. They report, however, that their results are not very sensitive to the precise definition.

[59] Rodgers, Brown, and Duncan report these components normalized as shares of the relevant error variance. For wage rates derived from reports of annual earnings and annual hours, the contribution due

Two studies focus on the accuracy of reports of starting wage in a particular job. Branden and Pergamit (1994) evaluated the consistency of respondents' reports of starting wages in the National Longitudinal Study by comparing responses reported at time *t* to those reported one year later. Only 42% of those studied reported the same starting wage for a particular job across the two years[60]. Consistency varied as a function of the time unit used for reporting, with higher rates of consistency among those reporting their starting wage as an hourly or daily rate (47% and 52% consistent, respectively) as compared to a consistency rate of approximately 13% for those reporting a biweekly wage rate. In contrast, Barron, Berger and Black (1997)[61] find a high correlation between employers' and employees' reports of starting wages (.974). Differences in the length of the recall period (one year vs. at most four weeks) most likely contributes to the differences in the findings from the two studies. Unfortunately, given these relatively short recall periods, neither study gives much evidence on the question of recall accuracy for starting wages of those who have been employed for longer periods (e.g., typical of information collected as part of a retrospective event-history question sequence).

On the whole, the evidence suggests that reporting of weekly or monthly earnings are highly correlated with employer reports. Available evidence on earnings per hour is much less reassuring. Unfortunately, the cautions from the various PSID Validation Studies are – as their authors indicate – likely to be overly dramatic because the true variance of hourly earnings is considerably smaller in one firm than in a broader sample.

As was true for annual earnings, a few of the studies in Table 1 attempt to assess the importance of measurement errors in frequently-estimated linear models. Mellow and Sider (1983) examined the impact of measurement error in wage equations; they concluded that the structure of the wage determination process model was unaffected by the use of respondent- or employer-based information, although the overall fit of the

to error in annual earnings and annual hours are about equal (.93 and .80). The errors are positively correlated ($r = .43$) and so the covariance is negative (–.74). For wage rates based on the most recent pay period, errors in reported earnings are about twice as important as errors in reported hours (1.36 and .62, respectively); the covariance is again negative (–.98). Based on usual pay, the contribution due to error in reports of earnings is 1.26, from error in reports of hours is .32, and the covariance is –.58.

[60] Only those who reported their pay in the same time unit in both interviews are included.

[61] The study reported by Barron, Berger, and Black was based on a sample of establishments with 100 or more employees, screened to determine whether they were hiring at the time of the initial interview. The data collection encompassed three interviews with the firm and three with the newly hired employee of the firm. Of the 5000 establishments originally sampled, no attempt was made to contact 1603 establishments due to budgetary restrictions. Of the 1554 establishments classified as eligible and for whom interviews were attempted, complete information was obtained from 258 (16.6%) employer–employee pairs. The low response rate, coupled with the lack of information for over 32% of the originally sampled establishments, raises serious concerns with the inferential limitations of the study. The authors report that the sample for which they could obtain information was similar to the original 5000 establishments in size and industry, but completions were more likely to come from rural areas and the Mountain and Pacific regions.

model was somewhat higher with employer-reported wage information. Bound, Brown, Duncan and Rodgers (1994) report estimates of simple "labor supply" equations (ln hours regressed on ln earnings per hour and demographic controls). Here, a number of potential biases are at work – due to the unreliability of hours reported as well as errors in hourly earnings – and their impact depends on the true supply elasticity. In the end, their results suggest such estimates may be badly biased, though the direction of the bias and the contribution of errors in measuring earnings per hour are less clear[62].

The studies reported in Table 1 provide conflicting indications of the relative accuracy of survey reports of monthly or weekly earnings, with some relatively old studies showing quite high correlations with record values. The calculation of hourly earnings appears to be most prone to error; the correlations between interview and record values are significantly lower for hourly earnings than for weekly, monthly, or annual earnings. In most of the studies, however, hourly earnings are calculated from separate reports of earnings and hours rather than based on direct reports of hourly earnings by respondents. The error in hourly earnings is therefore a function not only of misreporting of earnings (annual, weekly, or monthly) but also a function of the reporting of hours worked, the later being subject to high levels of response error (see Section 6.4). An empirical investigation that has not been reported to date is the comparison of the accuracy of direct reports of hourly earnings by household respondents with the hourly earnings reports calculated from reports of earnings and hours.

6.2. Transfer program income

Transfer program income can be categorized broadly as falling within one of two categories: relatively consistent recipiency status and income levels once eligibility has been established, and highly volatile recipiency status as well as income. As with most other episodic events, we expect that relatively stable behavioral experiences will be reported relatively accurately whereas complex behavioral experience (e.g., month to month changes in the receipt of AFDC transfer income) would be subject to high levels of response error. Respondents experiencing complex patterns of on/off recipiency status will most likely err on the side of failing to recall exceptions to the rule (e.g., the two months out of the year in which they were not covered by a particular program).

[62] French (1998) uses the PSID-VS data to correct estimates of the inter-temporal labor supply elasticity for measurement error. Within the context of his model, the covariance of the change in hours and the once lagged change in wages scaled by the variance in the transitory component of wages should give an estimate of the inter-temporal labor supply elasticity. The covariance terms involve covariances between current and twice lagged hours and wages. French allows for individuals to under-report the transitory component of wages and the transitory component of hours caused by the transitory component of wages to be under-reported, and for errors in wages and hours to be correlated, but otherwise that measurement error is classical. With these assumptions, French is able to use the PSID-VS to correct for measurement error. His results suggest that measurement error can not explain the failure of inter-temporal labor supply effects to explain short term movements in hours.

Depending upon the usual status quo for these respondents (receipt or nonreceipt), both under- and overreporting may be evident. In addition, for some transfer program income subject to social desirability bias, we would hypothesize that respondents would err on the side of underreporting receipt. Finally, misunderstanding as to the exact type of transfer program income received by the respondent may lead to the misidentification of recipiency, leading to underreporting of one type of income receipt and a corresponding overreport of another type of income receipt.

For most surveys, the reporting of transfer program income is a two-stage process in which respondents first report recipiency (or not) of a particular form of income and then, among those who report recipiency, the amount of the income. One of the shortcomings of many studies which assess response error associated with transfer program income is the design of the study, in which the sample for the study is drawn from those known to be participants in the program. Responses elicited from respondents are then verified with administrative data. As noted earlier, retrospective or reverse record check studies limit the assessment of response error, with respect to recipiency, to determining the rate of underreporting; prospective or forward record check studies which only verify positive recipiency responses are similarly flawed since by design they limit the assessment of response error only to overreports. In contrast, a "full" design permits the verification of both positive and negative recipiency responses and includes in the sample a full array of respondents. Validation studies which sample from the general population and link all respondents, regardless of response, to the administrative record of interest, represent full study designs. These would include the studies by Bancroft (1940), Oberheu and Ono (1975), Halsey (1978), Hoaglin (1978), and the more recent studies by Marquis and Moore (1990), Grondin and Michaud (1994), Dibbs, Hale, Loverock and Michaud (1995), Moore, Marquis and Bogen (1996), and Yen and Nelson (1996). The findings from the other studies cited in Table 2, many of which indicate a preponderance for underreporting by respondents with respect to receipt of a particular type of income, are to some extent an artifact of the study design. Rather than interpret the findings from these studies as indicative of a consistent underreporting bias on the part of the respondents, a more conservative conclusion may be to view the findings as illustrative of the types and magnitude of errors recipients can make with respect to program receipt.

There are several different ways of summarizing the frequency of reporting errors, which can give very different impressions of the accuracy of the data. One is the fraction of cases for which interview and record data disagree. Another is the difference between the fraction reporting receipt in the interview data and the corresponding proportion according to the records, which is the extent of net under- or overreporting. A third is the pair of conditional probabilities, π_{01} = Prob (interview = no | record = yes) and π_{10} = Prob (interview = yes | record = no) that determines the extent of bias when recipiency is used as a variable in a regression (Section 2.5).

These three measures are related: the probability of disagreement = $\pi\pi_{01} + (1 - \pi)\pi_{10}$, and net underreporting = $\pi\pi_{01} - (1 - \pi)\pi_{10}$. The probability of disagreement tends to be lower for programs with low true participation rates as long as $\pi_{01} > \pi_{10}$;

Table 2
Assessment of measurement error: transfer program income

Reference	Variables of interest	Validation source	Findings
Bancroft (1940, Table 1)	Public relief	Administrative records	Pr(interview = currently receiving | record = currently receiving) = .92; Pr(interview = never received | record = never received) = .84; (this biased down because records miss receipt > 2.5 years prior to interview)

David (1962)	Public assistance	Administrative records	7% of recipients (according to records) reported not receiving public assistance;
			Mean (interview) = \$2334;
			Mean (record) = \$2758;
			Mean (error) = –\$424, or 18% of record;
			r(interview, record) = .30;
			Errors unrelated to respondent race, sex, age (but N = 46)

Haber (1966, Tables 1, 3)	Social Security income of "beneficiary unit" (couple or non-married individual)	Administrative records	Mean (interview) = \$991;
			Mean (record) = \$1052;
			Mean (error) = –\$51, or 5% of record (s.e. = \$5);
			Underreporting greatest for youngest (62–64) respondents, the institutionalized, and those with high (true) benefit levels

Livingston (1969)[1]	Several types of "public assistance" reported in special census	Administrative records in Dane County (WI)	22% of known recipients failed to report receipt; Over 50% of those reporting receipt in census could not be matched to an administrative record; Among those who report assistance and receipt is corroborated in records: median reported in inteview is 73% of median in records. For old age assistance and AFDC separately, corresponding ratios are 80 and 70%, respectively
Hu (1971)	Cash and medical assistance	Administrative records	27% of recipients failed to report assistance receipt

continued on next page

Table 2, *continued*

Reference	Variables of interest	Validation source	Findings
Oberheu and Ono (1975)	AFDC participation “last month”, annual AFDC, and Food Stamp receipt “last month” among low-income households with children	Administrative records	Reporting of AFDC recipient for last month: Pr(interview = yes | record = yes) = .68; Pr(interview = no | record = no) = .77; Net underreporting of receipt = 2%; Among those who correctly report having received benefits, mean underreport = $89/month; Similar findings for AFDC receipt for last year; Reporting of Food Stamp participation for last month: Pr(interview = yes | record = yes) = .70; Pr(interview = no | record = no) = .85; Net overreporting of receipt = 6%
Vaughan and Yuskavage (1976)	Social Security income (CPS)	Administrative records	Among cases where both record and interview showed a positive amount received: 54% of interviews exceed record amount; 39% of interviews are less than record amount; 7% of cases agree within $10; Mean error = $68 (5% of average benefit); Mean absolute error = $225 (15% of average benefit)
Halsey (1978)[1]	AFDC, Unemployment Insurance	Administrative records	Among cases where record and/or interview showed a positive amount received: Mean amount of AFCD underreported by 25–30%; Mean amount of Unemployment Insurance underreported by 50%; *r*(interview, record) are in .40–.60 range
Hoaglin (1978)[1]	Social Security income, Supplementary Security Income, “welfare”	Administrative records	Median response error is $0 for “welfare” (combines AFDC, general assistance and other programs), SSI, and unemployment insurance reports; slightly negative for reports of monthly Social Security amounts

continued on next page

Table 2, *continued*

Reference	Variables of interest	Validation source	Findings
Vaughan (1978)	Social Security income	Administrative records	Pr(interview = yes $\mid$ record = yes) = .87; Most of the remaining 13% appear to misreport SSI income as Social Security rather than failing to report any transfer income at all
Klein and Vaughan (1980)[1]	AFDC receipt	Administrative records	Pr(interview = yes $\mid$ record = yes) = .86
Goodreau, Oberheu and Vaughan (1984, Tables 1,3,4)	AFDC receipt	Administrative records in California, North Carolina, Pennsylvania and Wisconsin	91% report receiving cash assistance, but only 78% correctly identify the payment of AFDC per se; Amount last month (those receiving any cash assistance): Mean (record): $286; Mean (household report): $276; Simple difference: $10 (3.5%); Reporting error negatively related to record amount; Among those receiving AFDC; 74% reported as AFDC; 13% reported as other transfers; 4% underreported by those reporting receipt; 9% received by those reporting no cash transfers
Marquis and Moore (1990)	AFDC, Food stamps, Unemployment Insurance Benefits, Workers Compensation, Social Security (OASDI), Supplemental Security Income and Veteran’s benefits as reported in SIPP	Administrative records in Florida, New York, Pennsylvania and Wisconsin	Underreporting by known recipients (A) (Pr(interview = no $\mid$ record = yes)) and relative net underreporting (B) (1 − Pr(interview = yes)/Pr(record = yes)) by program: Program — A — B AFDC — .49 — −.39 Unemployment insurance — .39 — −.20 Food Stamps — .23 — −.13 Supp. Security Income — .23 — −.12 Veterans′ benefits — .17 — −.03 Social Security — .05 — +.01

continued on next page

Table 2, *continued*

Reference	Variables of interest	Validation source	Findings

Grondin and Michaud (1994)	Unemployment benefits	Canadian tax returns	4–6% of reports on recipiency are in error;
			Net underreporting rate 3–4% for recipiency;
			Pr(interview = no | record = yes) = .11–.16;
			Pr(interview = yes | record = no) = .005–.011;
			Discrepancy between record and interview report exceeds 5% for approximately a third of those with non-zero amounts for both

Dibbs, Hale, Loverock and Michaud (1995)[1]	Unemployment benefits	Tax returns	Mean (record) = \$5600 Mean (interview) = \$5300 Mean (error) = \$300, or 5%
Moore, Marquis and Bogen (1996)[1]	AFDC, Food Stamps, Unemployment Insurance, and Supplemental Security Income as reported in SIPP; two experimental questionnaires	Administrative records in Wisconsin	Underreporting by known recipients (A) (Pr(interview = no | record = yes)) and overreporting by non-recipients (B) (Pr(interview = yes) | Pr(record = no)), by program: Program — A — B AFDC — .10–.12 — .03–.04 Unemployment insurance — .41–.44 — .01 Food Stamps — .12–.17 — .02–.03 Supp. Security Income — .08–.13 — .03 70% to 80% report AFDC, Food Stamps, and SSI within 5% of record; 20 to 30% accurately report unemployment insurance
Yen and Nelson (1996)[1]	AFDC	Administrative records in Washington	93% of the 49 000 eligible person-months reported correctly; Survey-based estimates of monthly participation exceeded record-based estimates of participation by approximately 1 percentage point

continued on next page

Table 2, *continued*

Reference	Variables of interest	Validation source	Findings
Bollinger and David (1997)	Food Stamp participation (SIPP) individual records aggregated to the household level	Administrative records	$\Pr(\text{interview} = \text{yes} \mid \text{record} = \text{yes}) = .88$ $\Pr(\text{interview} = \text{yes} \mid \text{record} = \text{no}) = .003$

[1] Unpublished paper, reported in Moore, Stinson and Welniak (2000).

relatively high values of both π_{01} and π_{10} can lead to near-zero net underreporting but imply significant biases in a regression context.

Focusing our attention first on reporting of receipt of a particular transfer program, among the full design studies, there does appear to be a tendency for respondents to underreport receipt, although there are also examples of overreporting recipiency status. For example, Oberheu and Ono (1975) report a low correspondence between administrative records and household report for receipt of AFDC (monthly and annual) and Food Stamps ($\pi_{10} \approx .2$, $\pi_{01} \approx .3$), but relatively low net rates of under- and overreporting[63]. In the study reported by Marquis and Moore (1990), respondents were asked to report recipiency status for eight months (in two successive waves of SIPP interviews). Although Marquis and Moore report a low error rate of approximately 1% to 2% (not shown in table), the error rate among true recipients is significant, in the direction of underreporting. For example, among those receiving AFDC, respondents failed to report receipt in 49% of the person-months. Underreporting rates were lowest among OASDI beneficiaries, for which approximately 5% of the person-months of recipiency were not reported by the household respondents. The mean rates of participation based on the two sources suggest little difference; absolute differences between the two sources differed by less than one percentage point for all income types. However, the rareness of some of these programs means that small absolute biases mask high rates of relative bias among true participants, ranging from +1% for OASDI recipiency to almost 40% for AFDC recipiency. In a follow-up study, Moore, Marquis and Bogen (1996) compared underreporting rates of known recipients to overreporting rates for known non-recipients and found underreporting rates to be much higher than the rate of false positives by non-recipients. They also note that underreporting on the part of known recipients tends to be due to failure to ever report receipt of a particular type of income rather than failure to report specific months of receipt.

In contrast, Yen and Nelson (1996) found a slight tendency among AFDC recipients to overreport receipt in any given month, such that estimates based on survey reports exceeded estimates based on records by approximately 1 percentage point. Oberheu and Ono (1975) also note a net overreporting for AFDC (annual) and Food Stamp recipiency (annual), 8% and 6%, respectively.

The studies vary in their conclusions with respect to the direction and magnitude of response error concerning the *amount* of the transfer, among those who report receiving it. Several studies report a significant underreporting of assistance amount [e.g., Livingston (1969), Oberheu and Ono (1975), Halsey (1978)] or significant differences between the survey and record reports [Grondin and Michaud (1994)]. Other studies report little to no difference in the amount based on the survey and record

[63] Oberheu and Ono's sample is restricted to low-income households. This is likely to lead to a larger value of π_{01} than would be obtained in samples with the full range of household incomes. For example, π_{01} would be increased by mis-reporting other transfers, and these would be more common in low-income households.

reports. Hoaglin (1978) finds no difference in median estimates for welfare amounts and only small negative differences in the median estimates for monthly Social Security income. Goodreau, Oberheu and Vaughan (1984) found that 65% of the respondents accurately report the amount of AFDC support; the survey report accounted for 96% of the actual amount of support. Although Halsey (1978) reported a net bias in the reporting of Unemployment Insurance amount of −50%, Dibbs, Hale, Loverock and Michaud (1995) conclude that the average household report of unemployment benefits differed from the average true value by approximately 5% ($300 on a base of $5600).

In general, studies that assess the accuracy of transfer data from household surveys do not provide analyses of how such errors affect the parameters of behavioral models. An exception is Bollinger and David (1997), who estimate a parsimonious model of response error from validation data and then combine this information into a model of Food Stamp participation using a broader sample. They find that estimated effects of wealth and predicted earnings are increased by such corrections, though they note that these results depend on the model of response error based on a relatively small validation sample ($N = 2685$, but with only 181 participants). They also note that low income households are much more likely to mis-report Food Stamp receipt because they confuse Food Stamps with other transfers they receive; high-income respondents do not have other transfer programs to confuse Food Stamps with. Thus, while many examples of differential measurement error in survey reports of transfers are due to deliberate under-reporting, Bollinger and David's example shows that differential errors may also occur inadvertantly.

Studies of receipt of transfer payments are often interested not only in which groups are receiving transfers and how much they receive at one point in time, but also in the duration of receipt, and so in the transitions into and out of recipiency. Marquis and Moore (1990) matched data from SIPP interviews to administrative records for major transfer programs. They find that the number of transitions (those starting to receive benefits, and those whose benefits end) are overstated by interview respondents for some benefit programs and understated for others. A more consistent pattern is that such transitions are over-stated when one compares the last month of the reference period of one interview with the first month of the next – the so-called "seam" – and understated when one compares reports for two months collected in the same interview[64].

Comparing the findings for transfers with those for earnings suggests several broad conclusions. First, there is evidence of under-reporting of transfers, in contrast to the approximately zero-mean errors we found for earnings, and such underreporting seems more important for AFDC and other public assistance than for Social Security. This is quite consistent with comparisons of aggregate estimates based on survey reports to

[64] The finding of more transitions at the "seam" than at other points in a retrospective history pieced together from a series of interviews has been documented repeatedly [Moore and Kasprzyk (1984), Burkhead and Coder (1985), Hill (1987)].

independent estimates of aggregate amounts received[65]. Second, both non-reporting and underreporting by those who report receiving positive transfer benefits contribute to this underreporting, though it is hard to draw firm conclusions about the relative importance of these two sources of error. Third, accuracy of reports for individual transfers is reduced by mis-classification; i.e., respondents who report receiving a transfer, and may even report the amount correctly, but incorrectly identify the program that provided the benefit. Fourth, the focus on extent of underreporting in most studies leaves us with very little evidence on the likely effects of errors in reporting transfers when benefits from individual programs are used as either dependent or explanatory variables in behavioral models[66].

6.3. Assets

The literature on accuracy of reports of individual assets (and so, implicitly, of net worth) is similar in important ways to the literature on transfer income. Comparisons of aggregate values based on survey reports to independent estimates of these aggregates suggests that underreporting is likely to be a problem [Curtin, Juster and Morgan (1989)][67]. The literature has therefore focused on the extent of such underreporting, rather than on the variance of the error relative to the variance of the true (record) value, or the correlation between errors and true values.

A limited number of studies have focused on the assessment of measurement error related to the reporting of assets and only one of these, the study by Grondin

[65] In 1990, CPS totals amounted to 97% of independently-estimated levels of Social Security and railroad retirement benefits, and 89% of Supplemental Security Income payments. In comparison, CPS captured only 72% of AFDC and 86% of other public assistance [U.S. Census Bureau (1993, Table C-1)].

[66] Since benefits received depend in part on choices made by the recipient, analysts often use some sort of instrumental variable procedure to account for this endogeneity; for example, the level of AFDC benefits available in a state might be used as an instrument for the reported benefit level. While one might hope that instrumenting would undo the bias from measurement error as well, we have stressed that this hope depends on the reporting error being "classical". Given that benefits are bounded (at zero) and zero benefits are in fact common, we suspect errors are likely to be mean-reverting. Particularly for programs such as AFDC where reporting seems least accurate, the effect of reporting error on the consistency of IV estimates deserves explicit discussion.

[67] Curtin, Juster and Morgan report that the 1983 Survey of Consumer Finances produces aggregate net worth estimates that are close to those based on external (flow-of-funds) sources. This "adding up", however, reflects a balance between substantial discrepancies on particular wealth components (e.g., SCF shows "too little" liquid assets but "too much" housing), and a close look at these discrepancies suggests that the external totals are often not very accurate benchmarks for the survey data (e.g., because of difficulties in the flow-of-funds accounts in separating household and business asset holdings). However, alternative wealth surveys show substantially lower levels of net worth than does SCF (PSID and SIPP being roughly 80 and 60% of SCF, respectively). Juster, Smith and Stafford (1999) report that wealth surveys conducted in the 1960s typically found about two thirds of the net wealth found in the external sources. CPS reports of interest and dividend *income* were 51 and 33% of NIPA totals [U.S. Census Bureau (1993)]. Thus, comparisons with external totals suggest that under-reporting is likely to be the norm, although failure to sample the wealthiest households also contributes to these discrepancies.

and Michaud (1994), focuses specifically on interest and dividend income generated from asset ownership. Several studies conducted during the 1960s examine the extent to which respondents accurately reported savings account and stock ownership, comparing survey reports with financial institution reports for a sample of respondents known to own the particular asset of interest. As noted above, reverse record check studies by design limit the detection of response error to underreports. Hence, one should be cautious in drawing conclusions concerning the direction of response error based on these studies. As noted in Table 3, between 5% and almost 50% of respondents fail to report existence of a savings account; 30% of those who own stock failed to report ownership. The high rate of underreporting is also evident in the full design validation study reported by Grondin and Michaud (1994).

Among those who report ownership of a savings account or stocks, the findings are mixed with respect to the accuracy of account amounts. Maynes (1965) and Ferber, Forsythe, Guthrie and Maynes (1969a) report a small amount of net bias for reports of savings account amounts (–5% and 0.1%, respectively), while Ferber, Forsythe, Guthrie and Maynes (1969b) report that 80% of respondents are accurate in their reports of stock holdings. In contrast, Ferber et al. (1969a) indicate that there is a large degree of response error, with only 40% of respondents reporting the account amount within 10% of the true value. Similarly, Lansing, Ginsburg and Braaten (1961) indicate an absolute discrepancy of almost 50% between financial records and household survey respondents' reports of saving account amounts, a discrepancy similar to that reported by Grondin and Michaud (1994).

A few studies attempt to validate survey responses to questions about the value of owner-occupied housing, a very important component of wealth for most households. Kish and Lansing (1954) find that owners' estimates are close to appraisers' on average (mean discrepancy = 4%) but the two estimates differ by 30% or more in a quarter of the cases. Scrutinizing cases with the largest discrepancies – which a typical survey, without validation data, could not do – they find that the largest discrepancies were due to coding errors (e.g., omitting a zero or misreading a lead digit in moving from the interview form to the data record). Rodgers and Herzog (1987) find that differences between household estimates of assessed value and property-tax records of assessed value are positively related to the record value. This contrasts with the negative correlation they find for other variables, and which is typically found in other studies.

Related perhaps to respondents' difficulty in providing accurate responses to questions about asset holdings is the substantial level of item non-response – it is not uncommon for 30% of those who report owning an asset to either refuse to provide or claim to not know the value of the asset [Juster and Smith (1997)]. In response, surveys have increasingly used "unfolding brackets": questions of the form "would it be more or less than X", where a "yes" ("no") to the first such followup leads to a second with a higher (lower) value of X. Thus, respondents unwilling or unable to provide a dollar amount are induced to specify a range in which they believe the value of their asset holding lies. Since those who are initially unwilling or unable to give a dollar

Table 3
Assessment of measurement error: assets[1]

Reference	Variables of interest	Validation source	Findings
Kish and Lansing (1954)	House value (1950 Survey of Consumer Finances)	Appraisals (conducted specifically for validation)	Mean value reported by owner = \$9560 Mean value reported by appraiser = \$9210 Difference = \$350 (s.e. = 170) Owner's and appraiser's estimates differed by at least 30% in 24% of the cases; Largest discrepancies ultimately traced to coding errors
Lansing, Ginsburg and Braaten (1961)[2]	Savings account ownership[3]	Financial institution records	Ownership of savings account: Pr(interview = yes | record = yes) = .75 Conflicting evidence on extent of underreporting by those who report having accounts: In one sample, mean (record) = \$3310, mean (error) = –500 or –14%; mean abs error = \$1571 In second sample, mean error = +2%
Maynes (1965)	Savings account ownership[4]	Financial institution records	Ownership of savings account: Pr(interview = yes | record = yes) = .95 Of those who report an account and the amount in it: Mean(record) = 1827, mean error = –83 or 5% Mean error = –1% for those who consult records, –10% for those who do not; Error negatively related to record amount; Savings over 9 months also underreported, with errors negatively related to record savings
Ferber (1966)[2]	Savings account ownership	Financial institution records	Ownership of savings accounts: in three samples, Pr(interview = yes | record = yes) = .81, .78, and .65 Among those reporting an account and the amount in it, mean errors = –20%, 0.3%, and 8% of record means

continued on next page

Table 3, *continued*

Reference	Variables of interest	Validation source	Findings
Ferber, Forsythe, Guthrie and Maynes (1969a, Tables 1–3)	Savings account ownership	Financial institution records	46% of known accounts not reported; 32% of families known to have at least one account reported not having any; Owners of larger accounts less likely to participate in survey, but more likely to report accounts if they participate; For accounts reported and matched to record data, mean error negligible (record = \$3040, interview = \$3042) but interview reports differ from record by ±50% for 28% of respondents; Reporting error negatively related to record value
Ferber, Forsythe, Guthrie and Maynes (1969b)	Stock ownership (in shares of particular cooperating firms)	Financial institution records	Ownership of stock (in a particular firm): Pr(interview = yes | record = yes) = .70 Those who own more shares less likely to participate in survey; among participants, reporting owning (any of) the stock not monotonically related to shares actually owned; For stocks reported and matched to record data, mean error negligible (record = 63.8 shares, interview = 63.9); interview reports differ from record by ±50% for 13% of respondents; Those with largest holdings tend to underreport, but otherwise relationship between error and record values is irregular
Rodgers and Herzog (1987)	Assessed value of house (Study of Michigan Generations)	Property tax records	*r*(error, record) = .242 (s.e. = .138) error uncorrelated with age, education, marital status, or race; correlation with income = .224 (s.e. = .123)

continued on next page

Table 3, *continued*

Reference	Variables of interest	Validation source	Findings
Grondin and Michaud (1994)	Asset ownership; interest and dividend income	Canadian tax returns	22% and 11% of survey respondents (in two studies, paper-and-pencil and computer assisted, resp.) misreport whether interest and dividend income was received; Net underreporting rate of 19 and 6 percentage points, resp. Pr(interview = yes | record = yes) = .58 and .78, resp. Pr(interview = no | record = no) = .98 and .96, resp. Of those with positive amount of income and dividend income in both interview and record, approx. 70% agree within 5% on the amount

[1] Most validation studies involving assets have focused on the respondent's ability to report the ownership of the asset and the amount of the asset rather than the income generated from the asset.

[2] Reported in Moore, Stinson and Welniak (2000).

[3] Two samples, one limited to accounts $\geqslant$ \$500, one limited to accounts $\geqslant$ \$1000.

[4] Sampling rate higher for large accounts; accounts <10 guilders excluded. Reported statistics unweighted. Data from Netherlands.

value for the asset tend to be those with higher true values[68], brackets help to reduce the underreporting typically found by comparing asset levels as reported in household surveys to external (aggregate) values. For example, Juster and Smith (1997, Table 8) report that bracket-based imputations produce 6–12% higher estimates of mean net worth than imputations not based on bracket information.

However, experiments in which the bracket boundaries are varied randomly find that the distribution of amounts that comes out of the brackets depends on the bracket boundaries themselves. For example, in one study the fraction of cases with savings accounts less than $10 000 was 49% with the first bracket question set X equal to $1000 but only 37% when the first bracket question set X = $20 000. This, in turn, has led to several attempts to obtain "corrected" estimates by jointly modeling the determinants of the asset value and the effect of the (randomized) bracket boundaries [e.g., Hurd et al. (1998), Hurd and Rodgers (1998)]. Both studies find responses are pulled toward the boundary in the first bracket question. Setting bracket boundaries with an eye toward maximizing the fraction of the variance in the asset that can be accounted for by the categorical responses will tend to place the first bracket boundary toward the middle of the distribution of the asset in question [Heeringa, Hill and Howell (1995)]. Consequently, it is likely that the error induced by "anchoring" effects is likely to be mean-reverting in most applications. Lacking validation data, however, it is hard to say much about the effects of using bracket-based imputed values in regressions that use wealth as either dependent or explanatory variable.

6.4. Hours worked

Obtaining validation data for workers' reports of how many hours they work per week has proved more difficult than obtaining earnings data. In general, the administrative records – income tax, unemployment insurance, and Social Security payroll tax – used in many of the studies in Table 1 include no comparable data on hours worked. The largest Federal establishment survey of payroll and hours collects hours only for production workers in manufacturing and non-supervisory workers in other industries. A Bureau of Labor Statistics study that considered obtaining hours information for all workers noted "Hours data are less available than total payroll for most categories of workers" [U.S. Bureau of Labor Statistics (1983, p. 22)].

While the number of empirical investigations concerning the quality of household reports of hours worked is limited, one finding consistently emerges. Regardless of whether the measure of interest is hours worked last week, annual work hours, usual hours worked, or hours associated with the previous or usual pay period, comparisons between company records and respondents' reports indicate that interview responses overestimate the number of hours worked. The findings from seven studies in which

[68] Hurst, Luoh and Stafford (1998) attribute to Donald Trump the observation that those who know how much their assets are worth can't be worth very much.

household reports of hours worked are compared to employer's records are reported in Table 4; all of these studies were also represented in Table 1. Findings from three studies in which the quality of the reports of hours worked is compared to time-use diary estimates are also reported in Table 4.

Carstensen and Woltman (1979) compared reports of "usual" hours worked per week. They found that compared to company reports, estimates of the mean usual hours worked were significantly overreported by household respondents, 38.4 hours vs. 37.1 hours, respectively, a difference on average of 1.33 hours, or 3.6% of the usual hours worked. Similarly, Mellow and Sider (1983) report that the mean difference between the natural logarithm of worker reported hours and the natural logarithm of employer reported hours was .039. They also report that the measurement error has a non-trivial variance (.064) but do not compare that variance to that of either the interview or the record hours variable.

In their reanalysis of this same data, Angrist and Krueger (1999) report a variance of the difference in ln hours of .083. This compares to the variance in ln survey hours of .195 or a signal to total variance ratio of roughly .8. Again, mean reversion will tend to reduce the implied attenuation to less than the .2 this number suggests.

Duncan and Hill (1985) find that worker reports of hours worked in the previous year (from the first wave of the PSID Validation Study) exceed company reports by 90 hours per year, nearly 6% of mean hours. The average absolute error was 157 hours. Recall of hours worked two years ago were less accurate, as expected, with a mean absolute error of 211 hours. More readily related to the discussion of biases in Section 2 is their finding that the ratio of error to record variance is .37. Bound, Brown, Duncan and Rodgers (1989) also find hours are overreported in the second wave of the Validation Study, though the mean error for ln (annual hours) is only .012, which is not statistically significant. However, the variance of the error is about .6 of the variance of record ln hours. Once again, there is evidence of significant mean reversion (correlation between error and true hours of −.37). Rodgers, Brown and Duncan (1993) consider various time intervals – hours worked in the previous year, hours worked in the previous pay period, and "usual" hours worked. They find the correlation between interview reports and company records is .61 to .66 for all three measures; and, for all three measures, the correlation between error and company records is −.31 to −.37. It is worth recalling that the PSID Validation Study obtained data from one manufacturing firm with few part-time workers and therefore, limited variation in hours per week, but (at least at the first wave) less than full-year employment for many workers. Moreover, hours data were unavailable for salaried workers. Barron, Berger and Black (1997) report a correlation between employers' records and respondents' reports of hours last week, .769; but this correlation falls to .61 for ln (hours).

One might wonder whether, in the case of hours, the company reported values should be treated as "true". For those who are paid by the hour, accurate recording of hours is essential for correctly paying the worker, and for those who "punch a clock" the company presumably has at least accurate records of the worker's coming and going.

Table 4
Assessment of measurement error: hours worked

Reference	Variables of interest	Validation source	Findings
Carstensen and Woltman (1979)	Usual hours worked per week (CPS special supplement)	Employers' records	Mean (household): 38.43 Mean (employer): 37.10 Mean (difference): 1.33 (s.e. = .10) Compared to company records, estimates of the mean "usual hours worked" significantly overreported by household respondents
Stafford and Duncan (1980)	Average work week (Time Use Study) and Hours worked last week (CPS)	Time-diary reports of various work activities	Mean (average hours/week): 41.8 Mean (time-diary reports): 36.8 [37.5 eliminating 11 outliers]; Change in mean work hours per week between 1965 and 1975 larger in time-use data than CPS reports of hours worked last week: CPS hours / Time diary Married men: –1.3 / –3.4 Married women: –0.5 / –7.8 [Time-use means exclude those working <10 hours/week]

Mellow and Sider (1983)[1]	Hours worked (CPS)	Employers' records	ln (worker report) – ln (employer report):
			mean = .039, variance = .064
			Regression model predicting difference indicates that professional and managerial workers report more hours than their employers; overreports also associated with educated and nonwhite employees, while females tend to underreport hours

continued on next page

Table 4, *continued*

Reference	Variables of interest	Validation source	Findings
Duncan and Hill (1985)[2]	Annual hours worked, year t and $t-1$ hourly workers only (PSID Validation Study)	Employers' records	Mean 1982 annual hours (interview) : 1693 Mean 1982 annual hours (record) : 1603 Mean difference : 90 ($p < .01$) Mean 1981 annual hours (interview) : 1880 Mean 1981 annual hours (record) : 1771 Mean difference : 115 ($p < .01$) Mean 1982–1981 simple change (interview) : –185 Mean 1982–1981 simple change (record) : –167 Mean difference : –17 Mean absolute |1982–1981| change (interview) : 357 Mean absolute |1982–1981| change (record) : 286 Mean difference : 70 Significant overreporting of hours worked for both years with an average absolute error of ≈ 10%; Error-to-record variance ratio: ln 1982 annual hours: .366
Hamermesh (1990)	Hours worked last week (Time use studies)	Time diary data	Average hours worked from CPS-like question on hours worked last week exceed time-diary estimates by 1.5 hours in 1975 and 3.6 hours in 1981
Rodgers, Brown and Duncan (1993)[2]	Hours worked: annual; most recent pay period; usual pay period. Hourly workers only (PSID Validation Study)	Employers' records	Correlation between self-report and company records: .66, .66, and .61 for annual, most recent, and usual pay periods, respectively, after deleting outliers. Relative ranking sensitive to this decision and so unclear, overall; Correlation between error and company records: –.31, –.36, and –.37 for annual, most recent, and usual pay periods, resp. Weak positive correlation (.061, not significant) of errors in annual earnings over time (1986 and 1982)

continued on next page

Table 4, *continued*

Reference	Variables of interest	Validation source	Findings
Robinson and Bostrom (1994)	Hours worked last week (Time Use Study)	Time diary data	Hours worked last week exceed time-diary estimates of hours worked per week by 1 hour in 1965, by 4 hours in 1975, and by 7 hours in 1985
Bound, Brown, Duncan and Rodgers (1994)[2]	Annual hours worked (PSID Validation Study)	Employers' reports	Mean interview reports of ln hours insignificantly higher than record values; A = ln annual hours 1986; B = 1986–1982 change in ln annual hours; A B variance (interview) .0180 .0620 variance (record) .0174 .0529 variance (error) .0104 .0237 *r*(interview, record) .7033 .7962 *r*(error, record) –.3701 –.2061 *b*(record on interview) .6828 .7355 Correlation between 1986 and 1982 errors positive but very small (.064)

Barron, Berger and Black (1997, Table 5.1)	Hours worked per week (Upjohn Institute Survey)	Employers' records	Mean (interview) = 38.5, Mean (record) = 37.0;
			difference in means statistically significant (t = 3.95);
			for ln (hours), difference in means = .031
			r(interview, record) = .769
			for ln (hours), *r*(interview, record) = .61

continued on next page

Table 4, *continued*

Reference	Variables of interest	Validation source	Findings
Angrist and Krueger (1999, Table 10)[1]	Hours worked (CPS)	Employers' records	ln (employee-reported hours) – ln (employer reported hours): mean = .043 variance (interview) .195 variance (record) .182 variance (difference) .083 *r*(interview, record) .780 *r*(error, record) –.149 *b*(record on interview) .870 Recoding lowest (highest) one percent of employee-reported wages to the 1st (99th) percentile value increased *b*(record on interview) to .91

[1] It is unclear from the empirical findings as to the time reference used for the reporting of hours worked.

[2] Sample limited to a single employer. Hours worked calculated from the respondents' account of each week of the year (working, sick or annual leave, etc.).

For other workers, the link between hours worked and pay is much less tight, and as noted above some employers may not even keep records of their hours.

Three of the studies represented in Table 4 take a different approach to assessing the worker's report of hours worked. In addition to CPS-like questions on hours worked per week, the time use studies obtained time diaries from respondents. These diaries involved detailed reporting of activities at each time in the previous day. While only a few days' time diaries were collected from each respondent, when aggregated across respondents, work hours reported in the time diaries should add up to those reported in CPS-like questions. All three studies in Table 4 that used time use data [Stafford and Duncan (1980), Hamermesh (1990), Robinson and Bostrom (1994)] report that CPS-style questions lead to higher estimates of work time than are obtained from the time diaries. The discrepancies are, if anything, larger than those between worker and employer reports, and the gap between CPS-like questions and the time-diary based estimates is growing over time.

Evidence on the importance of measurement error in interview-based measures of change in hours is available in the studies based on the PSID-VS. Duncan and Hill (1985) find that constructing the change in annual hours by differencing reports for two previous years asked in a single interview (the first PSID-VS wave) leads to a relatively noisy measure, with an error to true variance ratio of .8. While sizeable errors in changes calculated in this way are likely to be reduced by a positive correlation between the errors in the two years [Rodgers, Brown and Duncan (1993) report that correlation as .36 in the second PSID-VS]. Bound, Brown, Duncan and Rodgers (1989) calculate the change in ln hours as one would in a longitudinal survey, as the difference between the logarithm of 1986 hours (reported in 1987) and 1982 hours (reported in 1983). Whether measured by the error to true variance ratio, the correlation between interview and record values, or the regression of record value on interview, the change in hours data are slightly more reliable than the levels data.

Examination of a model with earnings as the left-hand-side variable and hours worked as one of the predictor variables indicates that the high correlation between the errors in reports of earnings and hours (ranging from .36 for annual measures to .54 for last pay period) seriously biases parameter estimates. For example, regressions of reported and company record ln annual earnings on record or reported ln hours, age, education, and tenure with the company provide a useful illustration of the consequences of measurement error. Based on respondent reports of earnings and hours, the coefficient for ln hours is less than 60% of the coefficient based on company records while the coefficient for age is 50% larger in the model based on respondent reports. In addition, the fit of the model based on respondent reports is less than half that of the fit based on company records (R^2 of .352 vs .780).

The small number of studies validating worker reports of work hours against employer reports provide little guidance on the relationship between errors in reporting hours and other variables. Mellow and Sider's (1983) regression explaining the difference between the two sources indicates that professional and managerial workers were more likely to overestimate their hours, as were respondents with higher levels of education and nonwhite respondents. In contrast, female respondents tended to underreport usual hours worked.

In contrast to the findings with respect to annual earnings, we see both a bias in the population estimates as well as bias in the individual reports of hours worked in the direction of overreporting. This finding persists across different approaches to measuring hours worked, regardless if the respondent is asked to report on hours worked last week (CPS) or account for the weeks worked last year, which are then converted to total hours worked during the year (PSID). The consistent direction of misreporting coupled with what appears to be a trend toward increasing discrepancy over time suggests that (1) respondents misinterpret the question (monthly CPS); (2) incorrectly account for weeks worked (March CPS supplement and PSID); or (3) overreport as a result of social desirability bias in the direction of wanting to appear to be working more than is true. The monthly CPS questions concerning hours worked ask the respondent to report the total number of hours worked, not hours spent at the employer's site or hours of paid work. One potential source of error may be a difference in the underlying concept of interest, with users of the CPS data examining hours of paid employment and respondents indicating total number of hours, regardless of location or pay. The approach used in the March CPS and PSID to obtaining hours worked requires that the respondent report the number of weeks worked in the previous year. The March CPS question even includes the word "about" suggesting that the respondent can provide a rough estimate of the number of weeks worked. Here we would speculate that once again, the bias is in the direction of errors of omission related to exceptions to the rule. That is, if the respondent has been fully employed during the previous year, short spell deviations will not be reported. Either approach to the collection of hours worked are subject to social desirability bias, if respondent's perceive the reporting of more hours as socially desirable.

6.5. Unemployment

Concern about the reliability of survey reports relating to unemployment focuses on a number of distinct but related issues. One question is how accurate are reports of current labor force status, in which individuals are classified as employed, unemployed, or not in the labor force. A related issue is how errors in reporting labor force status in one month affect estimates of various labor force transitions (e.g., leaving unemployment by finding a job or leaving the labor force), and estimates of the duration of spells of unemployment that are calculated from such transitions. Other studies have focused on the accuracy of retrospective reports, including the number of spells of unemployment, the duration of such spells (including on-going spells) and the total length of time unemployed in a particular period. Unlike the variables considered in previous sections, there are no employer or administrative records that allow one to verify whether non-working individuals are unemployed or not in the labor force.

6.5.1. Current labor force status, and transitions to and from unemployment

The most widely used data on current employment status come from the Current Population Survey, which asks a series of questions each month and on the basis of the responses classifies individuals as employed, unemployed (roughly, looking for work), or not in the labor force (not working and not seeking work)[69]. Correctly classifying individuals involves taking proper – according to official definitions – account of complications such as wanting to work but believing none is available, search for a new job while on paid vacation from another job, school teachers on summer vacation, etc. Concerned with the accuracy of these responses, CPS regularly re-interviews a subsample of its respondents, re-asking the standard questions (about the reference week covered by the original interview) and attempting to reconcile any differences that the re-interview uncovers.

Several of the studies in Table 5 report estimates of the probability that an individual initially classified as unemployed (or employed, or not in the labor force) will be judged as unemployed following the re-interview process. A consistent finding of these studies [Poterba and Summers (1984, 1986), Abowd and Zellner (1985), Chua and Fuller (1987)] is that 11–16% of those classified as unemployed are likely to be misclassified, with most of the re-classifications being to not in the labor force rather than to employed[70].

A distinct but related problem with the classification of labor market status is that those in households that are interviewed by CPS for the first time are more likely to be classified as unemployed than they are in later months. There is also a weaker tendency for fewer of those in their sixth and seventh months to be counted as unemployed[71].

[69] The CPS is collected each month from a probability sample of approximately 50 000 households; interviews are conducted during the week of the month containing the 19th day of the month and respondents are questioned concerning labor force status for the previous week, Sunday through Saturday, which includes the 12th of the month. In this way, the data are considered the respondent's current employment status, with a fixed reference period for all respondents, regardless of which day of the week they are interviewed. The design is a rotating panel design in which households selected for participation are interviewed for four consecutive months, followed by eight months of no interviews, and then interviewed for the same four months one year later. In any one month, 1/8 of the sample is being interviewed for the first time, 1/8 for the second time, etc.

[70] Poterba and Summers, and Abowd and Zellner take the reconciled status from the re-interview as the "true" status. Chua and Fuller note that *initial* reports on the re-interview survey are more consistent with the initial CPS interview for the fraction of the sample where reconciliation is carried out than on the fraction where it is not. This suggests that, contrary to instructions, those conducting the re-interviews are aware of the initial CPS response before the respondent has answered the initial re-interview status. Poterba and Summers speculate that knowing the initial report leads re-interviewers to minimize discrepancies (and hence the work required to reconcile them). This would imply the reconciled responses are biased toward the original reports, and so taking them as true leads to underestimate the extent of error in the regular CPS.

[71] Bailar (1975) reports that in 1968–69, the number counted as unemployed was 20% higher for those in their first month than the average regardless of month. This fell to 9% in 1970–72 [Bailar (1975)] and 8% in 1974–83 [Solon (1986)]. Over these same time periods, the number counted as unemployed is 5–7% lower in month 7 than overall, and 3–4% lower in month 6.

Table 5
Assessment of measurement error: unemployment

Reference	Variables of interest	Validation source	Findings
Morgenstern and Barrett (1974)[1]	Annual number of person years of unemployment (1964–71) (CPS; WES from March Supplement)	None; comparison of estimate based on annual report (WES) vs. recall for previous week (CPS)	Average percentage discrepancies between CPS and WES (standard errors in parentheses): White males : 3.25 (6.23) Black males : 4.32 (6.45) White females : 23.95 (6.45) Black females : 21.56 (7.43) WES (annual recall) tends to underestimate unemployment, with the greatest discrepancy for women and youths. In high unemployment years, tendency for WES to overstate the amount of unemployment; Some indication of overreporting of unemployment among males age 25–44 and females age 45 and older
Horvath (1982)[1]	Average estimate of weekly unemployment (CPS; WES from March Supplement)	None; comparison of annual unemployment data with average computed from monthly CPS	Underestimate of unemployment based on annual WES measure ranged from about 9 to 25%; underestimate smallest for periods of increasing unemployment; Unemployment during the first six months of the year less likely to be reported in WES than unemployment in second six months of the year

continued on next page

Table 5, *continued*

<table>
<tr><th>Reference</th><th>Variables of interest</th><th>Validation source</th><th>Findings</th></tr>
<tr><td>Bowers and Horvath (1984)[1,2]</td><td>Duration of unemployment spell (CPS)</td><td>None; reporting of continuous unemployment spell one month later compared to report at time <i>t</i></td><td>Approximately 25% of respondents consistent in their report of unemployment duration;
Percent consistent in reports of unemployment duration as a function of spell duration:
<table>
<tr><td>&lt; 5 weeks :</td><td>32.8%–40.0%</td></tr>
<tr><td>5–10 weeks :</td><td>23.3%–28.3%</td></tr>
<tr><td>11–14 weeks :</td><td>16.0%–21.2%</td></tr>
<tr><td>15–26 weeks :</td><td>18.6%–29.8%</td></tr>
<tr><td>27–51 weeks :</td><td>20.0%–23.1%</td></tr>
<tr><td>52 weeks :</td><td>0.0%–10.7%</td></tr>
<tr><td>53–99 weeks :</td><td>8.7%–18.6%</td></tr>
</table>
The longer the spell reported at time <i>t</i>, the smaller the increase in reported duration one month later</td></tr>
<tr><td>Poterba and Summers (1984, Table 1)</td><td>Employment status (CPS, May 1976)</td><td>CPS Reinterview Survey (after reconciliation with initial reports)</td><td>Pr(CPS report | Re-interview Status), May 1976; A = reinterview after reconciliation; B = initial CPS interview
<table>
<tr><td>A</td><td>B(employed)</td><td>B(unemployed)</td><td>B(not in LF)</td></tr>
<tr><td>Employed</td><td>.991</td><td>.002</td><td>.008</td></tr>
<tr><td>Unemployed</td><td>.036</td><td>.860</td><td>.104</td></tr>
<tr><td>Not in LF</td><td>.005</td><td>.003</td><td>.992</td></tr>
</table></td></tr>
<tr><td>Poterba and Summers (1984, Table 2)</td><td>Unemployment duration (CPS, June 1996)</td><td>Consistency with May 1976 interview</td><td>Only 32% of June duration reports were “consistent” with (i.e., 3–5 weeks greater than) May report. Inconsistent reports about evenly divided between those with June–May difference greater than 5 weeks and those less than 3 weeks;
Difference between reports tended to be too large for those who reported being unemployed <20 weeks in May, and too small for those unemployed >20 weeks in May.</td></tr>
</table>

continued on next page

Table 5, *continued*

<table>
<tr><th>Reference</th><th>Variables of interest</th><th>Validation source</th><th>Findings</th></tr>
<tr><td>Abowd and Zellner (1985, Tables 6,7)</td><td>Employment status (CPS, 1987–1982)</td><td>CPS Reinterview Survey (after reconciliation with initial reports)</td><td>Pr(CPS report | Re-interview Status), 1977–1982; A = reinterview after reconciliation; B = initial CPS interview
<table>
<tr><th>A</th><th>B(employed)</th><th>B(unemployed)</th><th>B(not in LF)</th></tr>
<tr><td>Employed</td><td>.988</td><td>.002</td><td>.010</td></tr>
<tr><td>Unemployed</td><td>.019</td><td>.886</td><td>.095</td></tr>
<tr><td>Not in LF</td><td>.005</td><td>.003</td><td>.992</td></tr>
</table>
Assuming classification errors are independent from one month to the next, correcting for such errors increases the fraction unemployed one month who remain unemployed the next from 54% to 64%, and the fraction moving from unemployment to not in the labor force falls from 21% to 13%</td></tr>
<tr><td>Akerlof and Yellen (1985)[1]</td><td>Average estimate of weekly unemployment (CPS; WES from March Supplement)</td><td>None; comparison of annual unemployment data with average computed from monthly CPS</td><td>Previous-year unemployment reports from the WES average 90% of those obtained from the monthly CPS for the same calendar year;
WES–CPS difference has grown more negative over time;
Underreporting on WES most severe for those under 25 and for women 25–54</td></tr>
</table>

continued on next page

Table 5, *continued*

<table>
<tr><th>Reference</th><th>Variables of interest</th><th>Validation source</th><th>Findings</th></tr>
<tr><td>Duncan and Hill (1985)[3]</td><td>Unemployment hours, year t (1982) and $t-1$ (PSID Validation Study)</td><td>Employers' records</td><td>Annual unemployment hours
<table>
<tr><th></th><th>1982</th><th>1981</th><th>1982–1981</th></tr>
<tr><td>Mean (interview)</td><td>169</td><td>39</td><td>131</td></tr>
<tr><td>Mean (record)</td><td>189</td><td>63</td><td>126</td></tr>
<tr><td>Mean (error)</td><td>–11</td><td>–16</td><td>5</td></tr>
<tr><td>Mean (absolute error)</td><td>52</td><td>45</td><td>77</td></tr>
</table>
Mean of the difference between interview and record data not significantly different from zero in either year;
Average absolute difference was large relative to average amount of unemployment in each year – about one-third the mean unemployment for 1982 and two-thirds for year 1981 (one- and two-year recall, resp.)</td></tr>
<tr><td>Poterba and Summers (1986, Tables II, V)</td><td>Employment status (CPS, 1977–1982)</td><td>CPS Reinterview Survey (after reconciliation with initial reports)</td><td>Pr(CPS report | Re-interview Status), 1981; A = reinterview after reconciliation; B = initial CPS interview
<table>
<tr><th>A</th><th>B(employed)</th><th>B(unemployed)</th><th>B(not in LF)</th></tr>
<tr><td>Employed</td><td>.982</td><td>.003</td><td>.015</td></tr>
<tr><td>Unemployed</td><td>.038</td><td>.887</td><td>.075</td></tr>
<tr><td>Not in LF</td><td>.024</td><td>.018</td><td>.958</td></tr>
</table>
Assuming classification errors are independent from one month to the next, correcting for such errors increases the fraction unemployed one month who remain unemployed the next from 54% to 73%, and the fraction moving from unemployment to not in the labor force falls from 21% to 9%</td></tr>
</table>

continued on next page

Table 5, *continued*

<table>
<tr><th>Reference</th><th>Variables of interest</th><th>Validation source</th><th>Findings</th></tr>
<tr><td>Chua and Fuller (1987, Tables 1–3, 6–7)</td><td>Employment status (CPS, 1976–1978)</td><td>CPS Reinterview Survey (<i>not</i> reconciled with initial reports)</td><td>Pr(CPS report | Re-interview report), 1976–1978; B = initial CPS interview
<table>
<tr><td>Re – interview</td><td>B(employed)</td><td>B(unemployed)</td><td>B(not in LF)</td></tr>
<tr><td>Employed</td><td>.967</td><td>.007</td><td>.026</td></tr>
<tr><td>Unemployed</td><td>.099</td><td>.661</td><td>.240</td></tr>
<tr><td>Not in LF</td><td>.042</td><td>.019</td><td>.939</td></tr>
</table>
Assuming classification errors are independent from one month to the next, correcting for such errors increases the fraction unemployed one month who remain unemployed the next from 48% to 67%, and the fraction moving from unemployment to not in the labor force falls from 21% to 7%</td></tr>
<tr><td>Mathiowetz and Duncan (1988)[4], Mathiowetz (1986)[4]</td><td>Unemployment spells (PSID Validation Study)</td><td>Employers' records</td><td>Overall, 66% of spells unreported;
Accurate reporting of spells associated with the amount of unemployment in a given month and the temporal complexity of the spell</td></tr>
<tr><td>Torelli and Trivellato (1989)[5]</td><td>Unemployment duration (youth 14–29 in Italy's quarterly labor force survey)</td><td>None; consistency in reporting duration of unemployment spell between quarterly surveys and actual elapsed duration</td><td>Around 40% of reports of unemployment duration are consistent; tendency to under- or overreport duration related to actual length of spell – the longer the duration of unemployment, the greater the propensity to underreport the duration. "Rounding" or "heaping" accounts for approximately one-third of the inconsistencies</td></tr>
</table>

continued on next page

Table 5, *continued*

Reference	Variables of interest	Validation source	Findings
Levine (1993)[1]	Unemployment rate (CPS and WES from March Supplement)	None; comparison of contemporaneous rate and one year retrospective recall	Unemployment rate underreported by 7% to 24% when comparing retrospective rate to contemporaneous rate; 35% to 60% of persons failed to report unemployment one year after the event. Misreporting rate related to length of unemployment spell; Error correlated with economic cycle – less underreporting during recessionary periods, greater underreporting during expansionary periods

[1] Estimates based on the monthly CPS unemployment rate used as the "gold" standard for comparison. Comparison involves the use of different questionnaires and different questions to obtain measure of unemployment.

[2] The authors note that the findings may, in part, reflect rounding by those with very long (>6 months) unemployment spells.

[3] Sample limited to a single employer. Hours unemployed calculated from the respondents accounting for each week of the year (working, sick or annual leave, etc.).

[4] Sample limited to a single employer. Spell-level information obtained by asking the respondent to report the month in which he or she was unemployed for at least part of the month. Company plagued by sporadic unemployment during years of interest.

[5] Sample limited to those ages 14–29 living in the Lombardy region of Italy.

This pattern of "rotation group bias" is also present in the re-interview data [Bailar (1975)], which serves as a reminder that the re-interview data should not be regarded as error-free.

If those who are mis-classified in one month are correctly classified (or misclassified in a different way) in the next month, the number of transitions from one state to another will be exaggerated. For example, some of those who appear to move from unemployed to not in the labor force will in fact have been out of the labor force in both months. The extent to which classification error in one month biases estimates of transitions between statuses depends on whether the errors are persistent or independent from one month to the next. Lacking direct evidence on this score, analysts assume that the errors in one month are unrelated to errors in the next. On this assumption, a significant fraction of the apparent transitions – in particular, .10–.18 of the roughly .5 probability of leaving unemployment from one month to the next – appear to be due to errors in classifying workers in each of the months; transitions from unemployment to not in the labor force are exaggerated more than are transitions from unemployment to employment.

While there has been significant effort devoted to gauging the likely effects of errors in measuring labor force status on transition rates, there is much less evidence on how such errors might affect analyses of the effects of various factors on such transitions. Poterba and Summers (1995) explore the consequences of errors in reporting employment status for estimates of the effects of unemployment insurance and welfare receipt on the probability of leaving unemployment. Initially, they model the reporting errors as fixed probabilities, independent of the explanatory variables. Correcting for reporting errors based on re-interview evidence has little effect on the estimated effects of unemployment insurance, but substantially increases the effect of welfare receipt on labor force withdrawal. They note, however, that previous work suggests reporting errors in one month are higher for those who were unemployed in the previous month. They present alternative estimates intended to capture this intuition, albeit informally. If these alternative estimates of the probability of classification error are correct, effects of unemployment insurance and welfare receipt on transitions out of unemployment are significantly underestimated due to such error.

6.5.2. Retrospective unemployment reports

A substantial number of studies have examined directly the quality of unemployment reports. These studies, reported in Table 5, encompass a variety of unemployment measures including annual number of person years of unemployment, weekly unemployment rate, occurrence and duration of specific unemployment spells, and total annual unemployment hours. Only one study reported in the literature, the PSID validation study [Duncan and Hill (1985), Mathiowetz (1986), Mathiowetz and Duncan (1988)], compares respondents' reports with validation data; the majority of the studies reported in Table 5 rely on comparisons of estimates based on alternative study designs or examine the consistency in reports of unemployment duration across rounds of data collection. In general, the findings suggest that retrospective reports of unemployment by household respondents underestimate unemployment, regardless of the unemployment measure of interest.

Several of the studies reported in Table 5 compare unemployment statistics based on reports to the monthly Current Population Survey (CPS) to those obtained from the Work Experience Survey (WES), a set of questions included in the March Supplement to the CPS. The studies by Morgenstern and Barrett (1974), Horvath (1982), and Levine (1993) compare the contemporaneous rate of unemployment as produced by the monthly CPS to the rate resulting from retrospective reporting of unemployment during the previous calendar year. The measures of interest vary from study to study; Morgenstern and Barrett focus on annual number of person years of unemployment, Horvath on average estimates of weekly unemployment, and Levine on the unemployment rate. Regardless of the measure of interest, the empirical findings from the three studies indicate that when compared to the contemporaneous measure, retrospective reports of labor force status result in an underestimate of the unemployment rate. The rate of underreporting, depending upon both the measure of interest, the population, and the year, ranged from as low as 3% to as high as 25%. The discrepancy between the retrospective WES and the contemporaneous reports is generally taken as evidence of recall error. Note, however, that the monthly status reports are based on a complex algorithm that combines answers to a series of questions, while the WES allows the respondent greater freedom in self-classifying.

Across the three studies, the underreporting rate is significant and appears to be related to demographic characteristics of the individual. For example, Morgenstern and Barrett (1974) report discrepancy rates of 3 to 24%, with the highest discrepancy rates among women (22% for black women; 24% for white women). Levine compared the contemporaneous and retrospective reports by age, race, and gender. He found the contemporaneous rates to be substantially higher relative to the retrospective reports for teenagers, regardless of race or sex, and for women. Across all of the years of the study, 1970–1988, the retrospective reports for white males, ages 20 to 59, were almost identical to the contemporaneous reports.

One of the strengths of these three studies is the ability to examine the underreporting rates across many years of data and the impact of economic cycle on the quality of the retrospective reports of unemployment. The findings suggest a relationship between economic cycle and the quality of retrospective reports; Morgenstern and Barrett's (1974) analyses indicate that in years of high unemployment, retrospective reports of unemployment from the WES *overstate* the amount of unemployment. Horvath (1982) found that in periods of increasing unemployment, discrepancies between estimates of the average weekly unemployment rate based on concurrent and retrospective reports were smaller than during other economic periods. Levine's (1993) findings were similar to those of Horvath; he found that underreporting declined during recessionary periods and increased during expansionary periods.

In contrast to the findings comparing the estimates of unemployment from CPS and the WES, Duncan and Hill (1985) found that the overall estimate of mean number of hours unemployed one and two years prior to the survey based on employee reports and company records did not differ significantly. However, micro-level discrepancies, reported as the average absolute difference between the two sources, were large relative to the average amount of unemployment in each year.

In addition to studies which examine rates of unemployment, person-years of unemployment, or annual hours of unemployment, several empirical investigations have focused on spell-level information, examining reports of the specific spell and duration of the spell. Using the same data as presented in Duncan and Hill (1985), Mathiowetz and Duncan (1988) found that at the spell level, respondents failed to report over 60% of the individual spells. Levine (1993) found that between 35% and 60% of persons failed to report an unemployment spell one year after the event. In both studies, failure to report a spell of unemployment was related, in part, to the length of the unemployment spell; short spells of unemployment were subject to higher rates of underreporting.

With respect to reporting the duration of a spell of unemployment, there is some evidence that the direction and magnitude of response error is a function of the length of the unemployment spell. For continuous spells of unemployment (that is, those that had begun in month x and which were ongoing at month $x + 1$) Bowers and Horvath (1984) compared reports of spell duration to the actual amount of time that had elapsed between the two interviews. They found, on average, that the increase in the reported duration of the unemployment spell exceeded the actual elapsed time between interviews. Torelli and Trivellato (1989) used a similar approach for a quarterly survey and found that approximately 40% of the reported spell durations were consistent with the actual elapsed time and that the magnitude of response error was a function of the actual length of the spell. Specifically, they found that the longer the duration of unemployment, the greater the propensity to underreport the duration. Approximately one-third of the inconsistent reports was attributed to rounding by the authors. Poterba and Summers (1984) also find that the increase in spell length between interviews is smaller for those with longer durations of unemployment.

The findings suggest that, similar to other types of discrete behaviors and events, the reporting of unemployment is subject to deterioration over time. The passage of time alone however may not be the fundamental factor affecting the quality of the reports. Some evidence suggests that the complexity of the behavioral experience is a significant factor affecting the quality of retrospective reports. Both the micro-level comparisons as well as the comparisons of population estimates suggest that behavioral complexity interferes with the respondent's ability to accurately report unemployment for distant recall periods. Hence we see greater underreporting among population subgroups who traditionally have looser ties to the labor force (teenagers, women). Although longer spells of unemployment were subject to lower levels of errors of omission, a finding that supports other empirical research with respect to the effects of salience, at least one study found that errors in reports of duration were negatively associated with the length of the spell. Whether this is indicative of an error in cognition or an indication of reluctance to report extremely long spells of unemployment (social desirability) is unresolved.

6.6. *Industry and occupation*

The measures discussed thus far are ones in which discrepancies between the gold standard, whether administrative records or reports obtained from preferred designs, have been attributed to the response process. In that process, the respondent, the interviewer, and the question wording as well as the content of the questionnaire can all contribute to that which is often labeled “response” error. Evaluation of error associated with the measurement of industry and occupation must consider yet another factor which could contribute to the overall quality of a measure, the error potentially introduced through the coding process. The literature on response error, however, contains little discussion of the extent to which coding (as well as other post data collection processing) contributes to the overall error associated with a particular measure, or specifically with the classification of industry and occupation. Therefore, in the discussion that follows, the reader is cautioned that although disagreement between household reported industry and occupation and administrative records is often classified as response error, coding/classification errors most likely contribute to the overall level of discrepancy.

Based on the small set of studies which have examined the quality of industry and occupation reports, the findings presented in Table 6 indicate that, in general, industry is reported more accurately than occupation. For both industry and occupation, not surprisingly, the agreement rate between employees’ and employers’ reports classified according to a single-digit coding scheme are higher than the resulting reports categorized according to the more detailed three-digit classification. Mellow and Sider (1983) report agreement rates between 84% and 92% for industry classification and between 58% and 81% for the classification of occupation (three-digit and one-digit classification schemes, respectively) in their Current Population Survey sample. Agreement rates are lower in the EOPP data, but Mellow and Sider indicate there is reason here to doubt the accuracy of the record report. Brown and Medoff (1996) compared industry classification of workers’ reports to the SIC codes for the employer, as listed by Dun and Bradstreet. Using fourteen industry groups, their comparison yielded an agreement rate of 79%. The findings from Mathiowetz (1992) are similar to those of Mellow and Sider, with occupational agreement rates of 52% to 76%, for three-digit and one-digit classifications, respectively. In the study by Mathiowetz, two sets of coders independently coded the reports of the employers and the employees while a third set of coders examined the two reports jointly to determine if the occupation could be considered the same occupation, that is, result in the same three-digit code. The direct comparison yielded an agreement rate of over 87%, suggesting that a significant proportion of the inconsistency in three-digit classification may be due to very subtle effects related to specific words used by the

Table 6
Assessment of measurement error: industry and occupation

Reference	Variables of interest	Validation source	Findings
Weiss, Dawis, England and Lofquist (1961)[1]	Occupation classification: 3-digit level	Employers' responses to independent questionnaire	Self-reports agreed with company reports for 67% of the jobs reported during the past five years; Agreement rate for current occupation 70%; for occupations more than 4 years ago, agreement rate drops to 60%; Agreement rate higher for older workers, but not related to education or broad occupation
Dreher (1977)[2]	Tenure (nine 5-year intervals)	Employers' records	r(interview, record) = .97
Mellow and Sider (1983)	Industry classification: 1- and 3-digit level (CPS and Employment Opportunity Pilot Project (EOPP))	Employers' reports	CPS (N = 4523): Agreement rate: 1-digit, 92.3%; 3-digit, 84.1%; Agreement rates only slightly higher for self- than for proxy reports; EOPP (N = 3327): Agreement rate: 1-digit, 87.5%; 3-digit, 71.1%
Mellow and Sider (1983)	Occupation classification: 1- and 3-digit level (CPS)	Employers' records	Agreement rate: 1-digit, 81.0%; 3-digit, 57.6%; Agreement rates only slightly higher for self- than for proxy reports
Mathiowetz (1992)[2]	Occupation classification: 1- and 3-digit level; direct comparison (PSID Validation Study)	Employers' records	Agreement rate: 1-digit, 75.7%; 3-digit, 51.8%; Direct comparison (coder looks at worker and employer description at same time) agreement rate: 87.3%
Brown and Medoff (1996)[3]	Industry classification: 14 industry groups	Dun and Bradstreet	Workers' reports and D&B SIC code agreed 79% of the time

[1] Sample limited to the first 325 persons of the Work Adjustment Project (Minneapolis–St. Paul metropolitan area) for whom both interview and employer work histories were obtained.

[2] Sample limited to a single employer.

[3] Sample limited to those respondents for whom respondents' reports of employer could be matched to D&B files. Successfully matched employers tended to be larger and in business longer than employers in the overall sample.

respondent to describe his or her occupation or used by the coder to classify the occupation.

For variables like industry or occupation with multiple classifications, the effect of measurement error on estimated parameters depends critically on the details of the discrepancies. For example, if those in high-wage industries misreport themselves to be in other high-wage industries, the bias in estimating industry wage effects will be less than if the misreporters are spread randomly across the remaining categories. Angrist and Krueger (1999, Table 11) calculate wage-weighted industry and occupation indices, based alternatively on worker and employer data, from Mellow and Sider's CPS sample. In univariate regressions, the effect of the industry index is biased downward by 8%, and occupation by 16%; controlling for standard covariates like education, potential experience, race, and sex leads to estimated biases of 10 and 25%, respectively. Hence, the general finding that occupation is measured "less accurately" than industry does seem to translate into larger biases, at least when the relative size of the coefficients associated with the various categories are constrained in this way.

With respect to the reporting of occupations, the evidence of the deleterious effect associated with longer recall periods is mixed. Weiss et al. (1961) report a decline in the agreement rate of occupational classification by employee's and employer's from 70% for the current occupation to 60% for occupations held more than four years prior to the date of the interview. Mathiowetz (1992) found no effect on length of recall period in the agreement between household and employer reports of occupation. Agreement rates between the two data sources for occupations held one year prior to the interview were 49% (3 digit) and 74% (1 digit) compared to 52% and 76%, respectively[72].

Given the difficulties in obtaining accurate measures of industry and occupation at one point in time, and the tendency of most workers to change industry and occupation only infrequently, there is general concern that measurement error will exaggerate the occurrence of changes in industry and occupation when estimates of such changes are obtained by comparing reports of industry and occupation obtained at two points in time. The extent of such exaggeration depends on the extent to which the measurement errors are independent (for a given individual) over time. Biases induced when change in industry or occupation is an explanatory variable will depend as well on the pattern of mis-classification (e.g., are those who in a high-wage industry but are mis-classified assigned to another high-wage industry). We have no direct evidence on the independence of such errors over time. Krueger and Summers (1988) assume an error rate for one-digit industries half as large as reported by Mellow and Sider (1983) (but with the same pattern of mis-classification as Mellow and Sider found),

[72] Because the respondents in the study by Mathiowetz were older, with more tenure, than a nationally representative sample, these estimates should be seen as conservative estimates of the decline in the quality of reporting occupation associated with an increase in the length of the recall period.

and assume such errors are independent over time[73]. They find such a correction has a more important effect on estimated industry wage differentials estimated from two successive years of CPS data (in which true changes are relatively infrequent, and so the fraction due to classification errors is probably high) than in data from the Displaced Worker Survey (where true changes are more common, and so the fraction due to classification errors is lower)[74]. The CPS results point clearly to the value of *evidence* on pattern of errors in measuring changes in industry and occupation.

6.7. Tenure, benefits, union coverage, size of establishment, and training

In addition to questions concerning earnings, hours employed or unemployed, and industry and occupation, many labor-related studies query the respondent as to their employment tenure, union membership, establishment or firm size, and the nature of various employment-related benefits. Few studies have investigated the quality of these various measures; Table 7 provides a summary of available findings. The lack of replication with respect to most of the measures of interests suggests that we err on the side of being conservative when drawing inferences from these studies.

We could locate only two studies of tenure with employer in which workers' reports are compared to employer records, and these are studies of individual firms. Agreement on starting date for current employer ranged from 71% [Weiss et al. (1961)] to over 90% in the the first wave of the PSID-VS [Duncan and Mathiowetz (1985)]; however, agreement in the former study was defined as a reported start date within one month of the company records and in the later study, as within one year of the actual start date. Bound et al. (1994) report that, in both PSID-VS waves, the correlation between interview and record data was .99[75]. In contrast, Brown and Light (1992) find that tenure reports in longitudinal surveys are often inconsistent – indeed, it is difficult to infer which survey years a worker was employed by the same employer from the tenure data alone. They consider a number of ways of resolving these inconsistencies (in tenure reports) and ambiguities (about when a worker has begun working for a new employer). Their main finding is that recoding the tenure values (so that tenure increases by the elapsed time between surveys if one infers that the worker has remained with the same employer) is important in applications where there are fixed effects for each worker or for each spell with a particular employer.

[73] Their taking half of Mellow and Sider's rate is intended as a rough correction for the fact that the employer reports in Mellow and Sider's include some errors, and for the fact that errors for an individual are probably not independent over time.

[74] For example, the wage difference between workers in manufacturing and otherwise similar workers in retail and wholesale trade is .07 in the CPS before correcting for measurement error, and .23 after correcting. For displaced workers, the difference changes from .11 to .13.

[75] Estimates of returns to tenure are about .002 higher (on a base of .01) in cross-section regressions when record rather than interview data are used. However, this difference is due to a correlation between tenure and errors in reporting earnings [Duncan and Hill (1985)].

Table 7
Assessment of measurement error: union coverage, tenure, firm/establishment size, and miscellaneous benefits

Reference	Variables of interest	Validation source	Findings
Weiss, Dawis, England and Lofquist (1961)[1]	Starting date	Employers' responses to independent questionnaire	Reported starting date agreed with company records (within one month) for 71% of the jobs in past 5 years; Validity significantly declines as a function of the length of time between start date and date of interview
Mellow and Sider (1983)	Coverage under union contract (CPS and Employment Opportunity Pilot Project (EOPP))[2]	Employers' reports	CPS (N = 4523) sample proportions (B = Employer report): Worker report B(covered) B(not covered) Covered .235 .030 Not covered .041 .694 EOPP (N = 1708) sample proportions: Worker report B(covered) B(not covered) Covered .362 .051 Not covered .098 .489
Duncan and Hill (1985)[3]	Coverage under union contract (PSID Validation Study)	Employers' records	Less than 1% disagreement, all in the direction of workers claiming coverage when employer did not
Duncan and Hill (1985)[3]	Union membership (PSID Validation Study)	Employers' records	Less than 1% disagreement
Duncan and Hill (1985)[3]	Health insurance, dental benefits, life insurance (PSID Validation Study)	Employers' records	Health Insurance: <1% disagreement; Dental Benefits: 5% disagreement (workers claim no benefits when employer indicates benefit); Life Insurance: 10% disagreement (workers claim no benefits when employer indicates benefit)
Duncan and Hill (1985)[3]	Paid time off for vacation days, sick days (PSID Validation Study)	Employers' records	Vacation: <1% disagreement; Sick Leave: 9% disagreement (3% claiming benefit when company record indicates no benefit; 6% employer claims benefit and employee reports no benefit)

continued on next page

Table 7, *continued*

Reference	Variables of interest	Validation source	Findings
Duncan and Mathiowetz (1985)[3]	Tenure (PSID Validation Study)	Employers' records	90% of respondents report hire date within one year of date recorded by company
Bound, Brown, Duncan and Rodgers (1994)[3]	Tenure (PSID Validation Study)	Employers' records	Correlation between worker and employer reports of tenure with employer = .99 in both 1982 and 1986
Brown and Medoff (1996)[4]	Establishment and company size	Dun & Bradstreet	Correlation between worker report and D&B value: .82 (ln establishment size) and .86 (ln company size)
Brown and Medoff (1996)[4]	Age of firm	Dun & Bradstreet	Correlation between worker report and D&B value: .56 (years firm in business) and .50 (ln years firm in business)
Barron, Berger and Black (1997, Table 5.1)	Union coverage	Employers' reports	Correlation between worker and employer report = .689
Barron, Berger and Black (1997, Table 5.1)	Eligibility for health and life insurance, and retirement plan	Employers' reports	Correlation between worker report and employer record: When first hired / After 2 years with firm Health insurance .590 .469 Life insurance .516 .508 Retirement plan .312 .327
Barron, Berger and Black (1997, Table 5.1)	Eligibility for paid vacation and sick pay	Employers' reports	Correlation between worker report and employer record: When first hired / After 2 years with firm Paid vacation .247 .490 Sick pay .294 .428
Barron, Berger and Black (1997, Table 5.1)	Hours of training	Employers' reports	Correlation between workers' reports and employers for various types of training: On – site formal training .398 Off – site formal training .457 Informal, managerial .176 Informal, coworker .379 Total training .475

continued on next page

Table 7, *continued*

<table>
<tr><th>Reference</th><th>Variables of interest</th><th>Validation source</th><th>Findings</th></tr>
<tr><td>Berger, Black and Scott (1998, Table 2)</td><td>Covered by employer-provided health insurance (March CPS)</td><td>Covered by employer-provided health insurance (April/May CPS Supplements)</td><td>March 1988 vs May 1988 CPS (N = 10 070) (B = March 1988 CPS):
<table>
<tr><th>May 1988 CPS</th><th>B(not covered)</th><th>B(covered)</th></tr>
<tr><td>Not covered</td><td>.094</td><td>.034</td></tr>
<tr><td>Covered</td><td>.073</td><td>.799</td></tr>
</table>
March 1993 vs April 1993 CPS (N = 11 603) (B = March 1993 CPS):
<table>
<tr><th>April 1993 CPS</th><th>B(not covered)</th><th>B(covered)</th></tr>
<tr><td>Not covered</td><td>.141</td><td>.040</td></tr>
<tr><td>Covered</td><td>.071</td><td>.748</td></tr>
</table>
</td></tr>
<tr><td>Berger, Black and Scott (1998, Table 3)</td><td>Eligible for health insurance (Upjohn Institute Survey)</td><td>Employers' reports</td><td>Employee vs employer reports (N = 257) (B = Workers' report):
<table>
<tr><th>May 1988 CPS</th><th>B(ineligible)</th><th>B(eligible)</th></tr>
<tr><td>Ineligible</td><td>.459</td><td>.156</td></tr>
<tr><td>Eligible</td><td>.054</td><td>.331</td></tr>
</table>
</td></tr>
</table>

[1] Sample limited to the first 325 persons of the Work Adjustment Project (Minneapolis–St. Paul metropolitan area) for whom both interview and employer work histories were obtained.

[2] In EOPP data, union coverage coded as yes if employer reports majority of (non-supervisory) workers are covered.

[3] Sample limited to a single company.

[4] Sample limited to those respondents for whom respondents' reports of employer could be matched to D&B files. Successfully matched employers tended to be larger and in business longer than employers in the overall sample. The authors note that due to potential inaccuracies in D&B counts of employer size, correlations listed above "probably understate the correlation between worker reports and perfectly accurate measures of employer size" (p. 280).

Several of the studies indicate that employees' reports of coverage under a union contract, union membership, insurance benefits, and vacation and sick leave are accurate. With respect to coverage under a union contract, Mellow and Sider (1983) report discrepancy rates of approximately 7% (CPS, national sample) and 15% (EOPP, national sample), and Barron, Berger and Black's (1997) estimate is 6%. This compares to a 1% disagreement rate reported by Duncan and Hill (1985) in their unionized single-employer sample. The EOPP disagreement rates are, however, inflated by the fact that the employer report is coded as covered or not depending on whenever a majority of the workers are covered. Freeman (1984) and Card (1996), using the Mellow–Sider CPS data but different sample definitions, find employers and workers disagreeing 3.5 and 5% of the time (respectively); like Mellow and Sider, they find the discrepancies about evenly divided between workers but not employers reporting coverage and the reverse[76].

In a simple bivariate regression of wages on union coverage, random zero-mean errors in measuring coverage would lead to an estimate whose proportional bias is equal to the sum of the two mis-classification rates ($\text{Prob}(x = 0 \mid x^* = 1) + \text{Prob}(x = 1 \mid x^* = 0)$). This bias is .19 and .31 in Mellow's CPS and EOPP data, respectively; .10 in Freeman's and .12 in Card's sample of Mellow and Sider's CPS data. Freeman stresses the extent to which this bias is inflated in longitudinal analyses. If measurement errors are independent over time, and the misclassification rates sum to .10, the bias becomes 29% in a fixed-effect model if 19% of the sample reports changing union status (as is true over 1970–79 in PSID); with smaller fractions of the sample changing coverage status (as would be true in studying one-year changes) the bias would be larger still. Of course, if the errors are positively correlated over time, the bias due to measurement error in a fixed-effect model would be smaller than under independence[77].

Card argues that, instead of treating the employer reports as "true", one should treat both employer and worker reports as subject to measurement error. He finds that the estimated impact of union status on wages is very similar using either worker or employer report, whereas the latter should have a larger effect if only the worker reports were subject to error. Indeed, he argues that both the wage equations and the patterns of agreement across industry are consistent with the hypothesis that both worker and employer reports are equally prone to error, with error rates (independent of true union status) of 2.5 to 3.0%. His model makes the common assumption that the error in reporting union status is uncorrelated with the error term in the wage equation. This rules out a number of plausible scenarios; for example workers who are not aware of

[76] Freeman also finds that, in two supplements to the May 1979 CPS (in which union coverage was asked in each), the responses given by workers are inconsistent 3.2% of the time. The inconsistencies were about equally distributed between those who said they were covered only on the first supplement and only on the second.

[77] Indeed, Freeman reports that misclassification rates summing to .10 would predict more changes of reported union status due to error alone than one observes in his 1974–1975 CPS panel.

being covered by a union contract being those in weak unions which fail to deliver high wages.

There appears to be considerable disagreement on the accuracy of employee reports of various fringe benefits. Duncan and Hill (1985) also report high levels of agreement for reports of health insurance (less than 1% disagreement), dental benefits (5% disagreement, all underreporting by the respondent), life insurance (10%, all in the form of underreporting by the respondent), number of vacation days (less than 1%), and number of sick days (9%, split between over- and underreporting). In contrast, Barron, Berger and Black (1997) report disagreement rates of 35% and 25% (with respect to initial benefits) and 19% and 13% (with respect to benefits after two years) for sick pay and life insurance, respectively. Berger, Black and Scott (1998) compare March CPS reports of employer-provided coverage to reports one or two months later in special CPS supplements. They find 11% of the reports are inconsistent, with lower overall coverage rates in the March surveys. Comparing employer and worker reports, they find that three fourths of the disagreements are workers who report they are eligible but whose employer reports them ineligible.

The focus of Mitchell's (1988) research was the respondent's knowledge of pension plan provisions. Using a match sample of household respondents and pension providers identified as part of the 1983 Survey of Consumer Finances, Mitchell finds pension misinformation as well as respondent's inability to answer questions concerning pension benefits to be quite widespread. The highest rates of inaccuracy by household respondents concerned knowledge of early retirement provisions; one third of the respondents could not answer the questions and among those who did respond, less than one third understood (or more specifically, could accurately report) the requirements for early retirement benefits.

These errors are likely to be particularly damaging in structural models that relate retirement decisions to pension incentives. As Gustman and Steinmeier (1999) note, workers with defined benefit pension plans do much better if they leave the firm at the early retirement age rather than even one year earlier. Thus, mis-reporting the age of early retirement eligibility by even one year can make it look like an individual is retiring at precisely the age at which economic incentives suggest retirement should not occur, and lead researchers to severely underestimate the importance of pension incentives. Gustman and Steinmeier also note that workers may in fact base their behavior on their perceptions rather than "true" incentives; for such workers, survey responses may be a better approximation for the variable that motivates behavior than is the "true" variable as calculated from the pension plans.

Using Dun and Bradstreet data as the record for comparison, Brown and Medoff (1996) examined the quality of household reports of establishment and company size as well as age of firm (i.e., how long the firm had been in business). Correlations ranged from .56 (correlation between worker report and D&B report of age of firm) to .82 (for ln establishment size) and .86 (ln company size). The authors note in their findings that potential inaccuracies in the Dun and Bradstreet records "probably understate the correlation between the worker reports and perfectly accurate measures of employer size".

Only one study reported in Table 7 examines the accuracy of respondents' reports of training hours. Barron, Berger and Black (1997) compared workers' and employers' reports of hours of training for several different types of training: formal training, training by co-workers, and training related to others performing the job. The correlation between the two reports was highest for off-site, formal training (.457) and for total number of hours of training (.457) and lowest for informal training by managers (.176).

The empirical research concerning the quality of respondent's reports of benefits, tenure, and industry characteristics is limited; in many cases we have only a single study to inform us as to the error properties of these measures. Although the findings suggest that the reporting of tenure and union coverage is highly correlated with administrative records, caution should be taken in drawing any conclusions from this limited literature. With respect to the reporting of fringe benefits, the findings are mixed. Based on the PSID-V, it appears that employees are well informed as to the characteristics of benefits whereas the studies by Barron, Berger and Black (1997) as well as Mitchell (1988) suggest high rates of inaccurate reporting.

6.8. Measurement error in household reports of health-related variables

While the examples discussed so far tend to be drawn from surveys that are most often used by economists, the empirical literature in several other substantive areas is rich with examples of the misreporting of autobiographical information. An important example is health, where work typically done by those in other fields provides evidence on the validity of health-related measures often used by economists and other social scientists.

6.8.1. Health care utilization, health insurance, and expenditures

As previously noted, much of the early work with respect to the assessment of the quality of retrospective reporting by survey respondents focused on the reporting of health care utilization, usually as reverse record check studies in which respondents were sampled from those with known hospitalizations or visits to physician offices. The design of these studies makes them well suited for investigating errors of omissions; however, many of these studies are uninformative with respect to overreporting errors.

Table 8 presents the findings from a selection of studies assessing either the reporting of health care utilization, characteristics of health insurance, or health care expenditures. Once again, we find evidence that response errors appear to be a function of the nature of the response task facing the individual, the length of the recall period, and the salience of the information to be retrieved.

Two of the studies reported in Table 8 assess the quality of reports of hospitalizations. Cannell, Fisher and Bakker (1965) describe a reverse record check

Table 8
Assessment of measurement error: health care utilization, expenditures, and insurance

Reference	Variables of interest	Validation source	Findings
Cannell and Fowler (1963)	Physician visits	Physician records	Errors of omission increase as a function of the length of recall: 15% unreported for one-week recall; 30% unreported for two-week recall
Cannell, Fisher and Bakker (1965)	Hospitalization utilization	Hospital records	Overall, 12% of hospitalizations not reported for one year recall period. Errors of omission related to: length of the recall period, ranging from 3% for hospitalizations within 10 weeks of interview, to 40% not reported for those occurring 52 weeks prior to interview; length of hospital stay, with longer stays (30+ days) subject to lower rates of omissions (~ 5%) than shorter stays (26% underreporting for stays of 1 day); perceived threat of the condition associated with the stay; 10% rate of omission for conditions judged to not be threatening, to 21% of those judged most threatening; Sample size of 1505 persons with 1833 hospital discharges during the past year.

continued on next page

Table 8, *continued*

Reference	Variables of interest	Validation source	Findings	
Yaffe and Shapiro (1979)	Hospitalizations, physician visits, dental visits, prescription medicines	Provider records (Note: agreement rates are for population estimates and not at the person level; data provided for two separate geographical areas)	Agreement rates for utilization:	
			Office-based physician visits	72%–83%
			Clinic visits	39%–54%
			Emergency room	94%–96%
			Dental visits	82%–86%
			Prescribed medicines	61%–75%
			Hospitalizations	94%–97%
			Agreement rates for expenditures:	
			Office-based physician visits	68%–78%
			Clinic visits	31%–38%
			Emergency room	65%–90%
			Dental visits	89%–99%
			Prescribed medicines	65%–77%
			Hospitalizations	87%–99%
			Based on completed interviews with 802 families with information for 2300 persons	

continued on next page

Table 8, *continued*

Reference	Variables of interest	Validation source	Findings	
Walden, Horgan and Cafferata (1982)	Health insurance characteristics	Health insurance plan information collected from insurance provider	Percent for whom agreement between two sources of information:	
			Existence of out-of-pocket payments	78%
			Amount of out-of-pocket payments	32%
			Sources of premium payments	74%
			Amount of premium paid by others	28%
			Agreement on coverage characteristics:	
			Semi-private hospital room	86%
			Physician in-patient surgery	88%
			Other in-patient physician	80%
			Maternity	55%
			Eye exam for glasses	73%
			Routine dental care	78%
			Orthodontia	69%
			Ambulatory x-rays; diagnostic tests	70%
			Ambulatory physician	54%
			Ambulatory prescriptions	54%
			Outpatient mental health care	29%
			Inpatient mental health care	32%
			Nursing home/similar facility	33%
			Based on data for 20 001 individuals	

continued on next page

Table 8, *continued*

Reference	Variables of interest	Validation source	Findings
Cohen and Carlson (1994)	Health care expenditures	Medical records	Expenditures by type of utilization (standard errors in parentheses):
			Inpatient hospital stays (*n* = 1050; 19% of hosptializations)
			Mean (household) $5228 (630)
			Mean (record) $4975 (451)
			Mean (simple diff.) $252 (451)
			Mean (absolute difference) $847 (210)
			Emergency room visits (*n* = 1765; 21% of emergency visits):
			Mean (household) $155 (10.7)
			Mean (record) $153 (7.1)
			Mean (simple diff.) $2 (8.9)
			Mean (absolute difference) $59 (9.0)
			Outpatient Department visits (*n* = 2609; 13% of outpatient visits):
			Mean (household) $227 (15.4)
			Mean (record) $238 (15.2)
			Mean (simple diff.) –$10 (17.1)
			Mean (absolute difference) $132 (17.0)
			Medical provider contacts (*n* = 17 169; 11% of medical provider visits):
			Mean (household) $47 (1.0)
			Mean (record) $53 (1.7)
			Mean (simple diff.) –$5.5 (1.6)
			Mean (absolute difference) $23 (1.5)

study[78] in which approximately 1500 respondents were asked to report on hospitalizations occurring during the previous 12 months. Overall, approximately 13% of hospitalizations were not reported. Response error, as measured by the percent of hospitalizations not reported by the respondent, increased as a function of the length of time between the date of the hospitalization and the date of the interview. For example, for hospitalizations occurring within 10 weeks of the interview, the underreporting rate was 3% whereas among hospitalizations occurring a year prior to the interview, 40% were unreported. The duration of the hospitalization was related to the rate of underreporting; 5% of longer hospital stays (e.g., those lasting 30 or more days) were unreported by the household respondent as compared to 26% of one-day stays.

Other studies have examined the quality of the reports related to utilization of office-based physician services. For example, Cannell and Fowler (1963) found that a significant proportion of office-based physician visits were unreported by the household respondent, even for recall periods as short as one week (15% unreported) and that the underreporting rate increased sharply with an increase in the reference period to two weeks (30% underreporting rate).

The Medical Economics Survey reported by Yaffe and Shapiro (1979) was designed to test the feasibility and effectiveness of several different survey design features to obtain information concerning health care utilization, expenditures, and health insurance coverage. The study included an assessment of face-to-face vs. telephone mode, as well as monthly vs. bimonthly interviews over a six month data collection period. In addition to the monthly or bimonthly interview, respondents were asked to maintain a diary (after the initial interview) to serve as a record-keeping system and memory aid for subsequent interviews. Prior to each follow-up interview and at the end of the study period, a cumulative summary of previously reported information was mailed to each household. Respondents were asked to review the report and to make any necessary additions or corrections, including entries about bills received since the time of the last interview. All medical care providers identified by the respondent as having provided care for anyone in the family during the study period as well as providers identified as the usual source of care, were contacted after the household data collection.

Several of the design features, specifically, the multiple rounds of data collection, coupled with the relatively short reference period, the use of a household diary, and the use of a summary were all included so as to minimize response error. These design features may account, in part, for the higher levels of agreement reported in Table 8 for this study as compared to other studies. In addition, Yaffe and Shapiro only report agreement rates for population estimates, that is, $(Y_{\mathrm{hh}}/Y_{\mathrm{med}})^{*}100$, where Y_{hh} represents the population estimate based on the household report and Y_{med} represents the population estimate based on the medical records. The estimates are provided for

[78] The sample consisted of persons selected from hospital records as well as a supplementary sample of persons without hospitalizations, so as to blind the interviewers as to the purpose of the study.

the two distinct geographical areas studied. As can be seen from the table, agreement rates for utilization are quite high for the most salient events (and less frequent) such as hospitalizations and emergency room visits, with agreement rates in excess of 90%. Agreement rates were lowest for clinic visits, 39 to 54%. With respect to expenditures, we once again see a high level of agreement between the two data sources for hospitalizations (87 to 99%) and the lowest agreement rates for hospital clinic visits (31 to 38%).

Cohen and Carlson (1994), using data from the National Medical Expenditure Survey, also examined the quality of household reports of medical expenditures. The entries in Table 8 present the mean household estimate, the mean medical record estimate, the mean of the simple difference and the mean of the absolute difference between household and medical record reports of total expenditures for each of four categories of utilization. The sample sizes provided in the table represent the number of events on which the estimates are made; the percent indicates what proportion of all household events of that type are included in the analysis. Due to the design of the NMES (which included a medical record component for a sample of all households) as well as provider nonresponse and inability to match events reported by the household with events abstracted from medical records, not all events reported by the household respondent were included in the analyses. In addition, the analysis is limited to those events for which there was expenditure data from both the household and medical record files. The comparison of the two data sources indicate that although the simple differences tend not to be statistically significant, the absolute differences clearly indicate significant disagreement between the two data sources.

Very few studies have examined the ability of household respondents to report detailed information concerning features of their health insurance. Knowledge of the existence of out-of-pocket payments and sources of premium payments was quite high (78% and 74%, respectively), but quite low with respect to amounts of out-of-pocket payments and amount of insurance premiums paid by others (less than 30% for each) [Walden, Horgan and Cafferata (1982)]. As we would expect, the majority of respondents were able to accurately report the standard major categories of coverage (hospital room, physician in-patient surgery, other in-patient physician services, and dental services). Knowledge of coverage associated with richer benefit plans was much lower, however, with less than one-third of the respondents correctly identifying whether or not their insurance covered outpatient mental health, in-patient mental health or nursing home services.

6.8.2. Health conditions and health/functional status

Measurement error in health surveys is not limited to the reporting of utilization, expenditures, and health insurance characteristics, but is also evident in the reporting of medical conditions as well as the reporting of health and functional status. Findings from a sampling of the literature which addresses the validity and reliability of self reports of health conditions and functional status are presented in Table 9.

Table 9
Assessment of measurement error: health conditions and health/functional status

Reference	Variables of interest	Validation source	Findings
National Center for Health Statistics (1961)	Chronic conditions	Medical records	Correspondence between medical records and household reports of chronic conditions ranged from 20% to 44%. Higher rates of correspondence between two data sources associated with: use of a checklist rather than free recall higher number of physician visits during the past year associated with the condition most recent physician service within the past two weeks No differences in the quality of the report by self/proxy status, age, gender, or race; Sample of approximately 1400 families; medical records indicate 4648 chronic conditions among respondents
National Center for Health Statistics (1967)	Chronic conditions	Medical records	Errors of omission as a function of time since last visit and response task: Time Recall Recognition < 2wks 58% 32% 2wk –4mo. 79% 51% > 4mo. 84% 66%
Katz, Downs, Cash and Grotz (1970)	Index of Activities of Daily Living; study of 270 patients at discharge	Correlation between index scores and other assessment scales	Correlation coefficient of .50 with mobility scale and .39 with house confinement measure
Madow (1973)	Chronic conditions	Medical records	46.8% of conditions recorded in medical records unreported in household interview (underreporting) while 40.4% of household reported conditions were not listed in medical record (overreporting); Interviews with approximately 5000 persons with over 15 000 conditions obtained from the two data sources

continued on next page

Table 9, *continued*

Reference	Variables of interest	Validation source	Findings
Bergner, Bobbitt, Carter and Gilson (1981)	Sickness Impact Profile (SIP); various trials of different samples	Test–retest reliability; internal consistency reliability (Cronbach's α)	Test–retest reliability scores ranged from .97 for interviewer-administered, to .87 for self-administered forms. Alpha coefficients for 136-item version = .94 (interviewer administered) and .81 (self administered)
Johnson and Sanchez (1993)	Health conditions	Medical records	Percent of medical events for which there is agreement between household and medical record report: 3-digit condition classification 40.4% 131-condition summary grouping 54.4% 20 category summary grouping 68.2% Based on 33 514 health events for which information was available from both the medical records and the household respondent.
McHorney, Ware, Lu and Sherbourne (1994)	Eight scales from the SF-36 instrument; study conducted among 3443 patients with one or more chronic conditions	Internal consistency reliability (Cronbach's α)	Physical functioning .93 Role physical .84 Bodily pain .82 General health .78 Vitality .87 Social functioning .85 Role emotional .83 Mental health .90

continued on next page

Table 9, *continued*

Reference	Variables of interest	Validation source	Findings	
McHorney, Kosinski and Ware (1994)	Eight scales from the SF-36 instrument; study conducted among national sample of adults, n = 1692; self-administered	Internal consistency reliability (Cronbach's α)	Physical functioning	.94
			Role physical	.89
			Bodily pain	.88
			General health	.83
			Vitality	.87
			Social functioning	.63
			Role emotional	.81
			Mental health	.82

In two reverse record check studies [National Center for Health Statistics (1961, 1967)], respondents were asked to report on the prevalence of chronic conditions. The second study also included an experiment designed to address the difference in the quality of data obtained from free recall as opposed to recognition from a checklist of conditions. The findings from these studies suggest that underreporting is a function not only of the length of the recall period (measured as the time since the last physician visit related to the condition), but also of the response task. Questions which frame the task as one of recognition as opposed to free recall resulted in lower rates of underreporting. However, for both response tasks, the underreporting rate was quite high, ranging from 32% underreporting for the recognition task related to the events occurring within the previous two weeks to an underreporting rate of 84% for free recall of events occurring four or more months prior to the interview. The improved reporting related to the recognition task is predictable; the presence of a cue provides both additional context for the respondent to understand the goal of the questions and an additional means for accessing the associated network of memory. The study by Madow (1973) is a complete record check design, limited to respondents in a specific health plan. As can be seen from the table, almost half of the conditions recorded in the medical records were not reported by the household respondent whereas over 40% of the conditions reported in the household interview were not identified in the medical record.

As part of the National Medical Care Expenditure Survey (NMES), Johnson and Sanchez (1993) examined the congruence between medical conditions as reported by the household respondent and medical conditions as reported by the medical care provider. These data are based on the same matched sample of household reported events and provider reported events used by Cohen and Carlson (1994) in their analyses of the quality of household reports of health care expenditures. Household reports reflect conditions associated with hospitalizations, visits to emergency rooms, outpatient departments, as well as office based physician visits. Household reported conditions, which reflect a mix of self and proxy collected information, were coded to three-digit level of detail by experienced coders using the International Classification of Diseases, Version 9 (ICD-9). ICD-9 condition codes were abstracted from the medical records, independent of the knowledge of the condition described by the respondent. Household reports of utilization were linked to the medical record abstracted records via a probabilistic match function. One of the variables used in the probabilistic match was a one-digit collapsed classification of the condition related to the utilization. As a result, the agreement rates – which indicate the percent of medical events reported by the household respondent for which the two condition codes (household based and medical record based) agree – are likely to be optimistic. At the three-digit level of detail, there is agreement between the condition codes as reported by the household and the medical condition recorded in the medical records for less than half of the medical events. As we would expect, grosser levels of aggregation result in higher rates of agreement.

While the lack of congruence between survey data and medical records is disturbing, we want to emphasize that this information alone tells us very little about the effect of this measurement error on parameters estimates. First, it seems plausible that reporting errors decline with the severity of the condition (severe arthritis is more likely to be reported than is mild arthritis). Second, in many cases researchers will be interested in modeling jointly effects of various conditions on outcomes. In such cases, it is hard to say much about either the magnitude or the direction of the bias on a single coefficient, since the coefficient on any one condition will be biased not only by the under and overreporting of that condition, but also by the under and over reporting of other conditions (see the discussion in Section 2.2).

Table 9 also examines the reliability, and to the extent possible, the validity of several measures of health and functional status. The measures examined include the Index of Activities of Daily Living [Katz, Ford, Moskowitz, Jacobsen and Jaffe (1963)], the Sickness Impact Profile [Bergner, Bobbitt, Kressel, Pollard, Gilson and Morris (1976)], and the SF-36 [Ware, Snow, Kosinski and Gandek (1993)]. In contrast to the validation studies presented earlier, no external measure of validity exists for the majority of the measures related to health or functional status. Rather, as with most psychometric scales, the interests lies in the reliability of the measure (that is, test–retest reliability or internal consistency) or the validity of the index, measured as the correlation or consistency with other subjective scales.

Despite its broad use, there has been little published with respect to the assessment of the validity or reliability of the Index of Activities of Daily Living, especially within the general population. Katz, Downs, Cash and Grotz (1970) applied the Index of ADLs as well as other indexes to a sample of patients discharged from hospitals for the chronically ill and report a correlation between the index and a mobility scale and a confinement measure of .50 and .39, respectively. Most assessments of the Index of ADL have examined the predictive validity of the index with respect to independent living [e.g., Katz and Akpom (1976)] or length of hospitalization and discharge to home or death [e.g., Ashberg (1987)]. These studies indicate relatively high levels of predictive validity.

Despite these findings, there is a growing body of literature that suggest that the measurement of functional limitations via the use of ADL scales is subject to substantial amounts of measurement error and that measurement error is a significant factor in the apparent improvement or decline in functional health observed in longitudinal data. For example, Mathiowetz and Lair (1994) found that conditions of the interview, characteristics of the interviewer, and type of respondent (self or proxy) were predictive of improvement in functional status over the 18 months of interest whereas the individual's demographic characteristics and health status were indicative of decline in functional status. Rodgers and Miller (1997) examined the consistency with which respondents reported functional limitations, using alternative sets of question wording. Consistent with other findings in the literature, they found that minor differences in the wording of questions resulted in significant variation in the proportion of respondents identified as being limited in one or more functional activities, ranging from a low of 6% (based on a single question) to more than 25% of the respondents[79] (based on a set of six to nine ADL questions).

The Sickness Impact Profile (SIP) measures health status by assessing the way sickness changes daily activities and behavior and consists of 136 statements grouped into twelve categories of activities. The profile focuses on actual performance as opposed to capacity. Bergner, Bobbitt, Carter and Gilson (1981) report on the reliability of the profile for both interviewer administered questionnaires and self-administered forms, with reliability higher for the interviewer administered form (.97) than for the self-administered form (.87). Internal consistency, as measured by Cronbach's alpha[80] was similarly lower for the self-administered form (.81) than for the interviewer-administered form (.94).

The SF-36 is a generic health status measure, one that is not specific to age, disease, or treatment, that focuses on health-related quality of life outcomes. The index covers eight areas of health: physical functioning, role limitations due to physical health problems, bodily pain, general health, vitality, social functioning, role limitations due to emotional problems, and mental health. The measure is designed for both interviewer administration as well as self-administration and both modes of data collection have been assessed with respect to validity and reliability. Reliability of the SF-36 has been assessed in numerous studies [see Ware et al. (1993) for summary of these studies]; across the various scales of the SF-36 and across the various studies, the median of the reliability coefficients equals or exceeds .80 (Cronbach's alpha). The findings from two of the more recent studies examining the SF-36 are reported in Table 9. McHorney, Ware, Lu and Sherbourne (1994) examined the internal consistency of the SF-36 among approximately 3500 patients with one or more chronic conditions; as can be seen from the table the coefficients range from .78 for general health to .90 for mental health. A self-administered version of the questionnaire study conducted among a nationally representative sample of noninstitutionalized adults found similarly high measures of internal consistency [McHorney, Kosinski and Ware (1994)].

6.9. Education

Despite the importance of schooling as both an outcome and as an explanatory variable in economic models, relatively little effort has been devoted to assessing the accuracy of survey reports of years of schooling or similar measures of educational attainment.

[79] Rodgers and Miller's study is based on the respondents to the first wave of the AHEAD study.

[80] Cronbach's alpha provides an estimate of internal-consistency reliability based on the average inter-item correlation and the number of items in the scale, expressed as $k\ r/[1 + (k - 1)r]$ where k equals the number of items in the scale and r is the average correlation between items. The coefficient alpha will be higher (1) the more questions asked about the topic, and (2) the higher the average correlation between the scores for all possible combinations of the entire set of questions. In most applied studies, the lowest acceptable level of internal consistency reliability is .70 for group data and .90 for individual-level analysis [Nunnally and Bernstein (1994)].

The literature which is available, however, illustrates a number of interesting issues that are potentially relevant for other variables as well.

Typically, these studies (summarized in Table 10) have two interview-based measures of education, each of which is plausibly measured with error. In assessing what we can learn from such data, recall that the OLS bias in estimating β in the model $y = \beta x^* + \epsilon$ when instead of x^* we use $x_1 = x^* + \mu_1$ depends on $1 - \lambda_1 = 1 - (\sigma_{x^*,x_1}/\sigma^2_{x_1})$. If we have another measure of x^*, $x_2 = x^* + \mu_2$ then $\lambda_1 = \sigma_{x_1,x_2}/\sigma^2_{x_1}$ as long as μ_2 is uncorrelated with x^* and with μ_1. In other words, as long as the error in measuring x_2 is "classical" whether x_2 is itself a particularly reliable indicator of x^* is unimportant. If, in contrast, μ_1 and μ_2 are positively correlated, the covariance between the two measures of x^* will overstate λ_1, and holding that correlation constant, the larger the measurement error in x_2 the worse the overstatement will be.

An early study of the reliability of reported years of schooling is Siegal and Hodge's (1968) analysis of 1960 Census data. Validation data came from the Post-Enumeration Survey (PES), a re-interview conducted to assess the accuracy of the original Census reports. They found that the Census reports and PES data on individual years of schooling are highly correlated. They also noted, however, that the variance of the Census report is slightly smaller than that of the PES education variable, which is inconsistent with the usual assumption that the Census report is equal to the true (PES) variable plus an uncorrelated measurement error. The discrepancy between the two reports was in fact negatively related to the PES value ($r = -.20$). They argued that one should expect errors to be negatively related to true values for bounded variables, since for those with the highest (lowest) true level of education, errors must be negative (positive). Given that the variances of the Census and PES variable are essentially equal, the $b_{\text{PES, Census}} = .93$, so the bias due to errors in measuring education as an explanatory variable is small (as long as other explanatory variables are not highly correlated with education).

Siegal and Hodge (1968) recognized the possibility that the PES measure of education is also measured with error and considered several relatively elaborate models in which both years of schooling and income are mis-measured. These relied on rather arbitrary identifying assumptions, and Siegal and Hodge concluded "we have not been able to devise an entirely plausible solution".

Bishop (1974) presents a comprehensive summary of the reliability of Census and CPS reports of education. Estimates of the correlation between Census and other measures of education center on .9, as do the alternative estimates of λ_1. Bishop notes that mean reversion would tend to reduce the bias caused by measurement error, while positive correlation in the errors would lead the values of λ_1 to be too high.

Bielby, Hauser and Featherman (1977) compare Current Population Survey reports to subsequent interviews and re-interviews of the same households approximately six to seven months later as part of the Occupational Change in a Generation (OCG) study. Focusing on the sample of non-black males that participated in both the OCG interview and re-interviews, they find inter-correlations among the three measures of years of schooling of .80–.92. The OCG shows both lower correlation with the other two

Table 10
Assessment of measurement error: education

Reference	Variables of interest	Validation source	Findings
Siegal and Hodge (1968, Table 2.1, Figure 2.1)	Years of schooling, 1960 Census	Post-enumeration Survey (PES)	variance (interview) = 12.82 variance (record) = 13.03 r(inteview, record) = .933 r(error, record) = –.205
Bishop (1974, Table 1)	Years of Schooling, Census, 1950–70	Current Population Survey (CPS); Post-enumeration Survey (PES); Census Reinterview Survey (CRS)	1970 Census r(Census, CPS) = .88 (males), .88 (females) b(CPS, Census) = .88 (males), .86 (females) b(Census, CPS) = .89 (males), .89 (females) 1960 Census r(Census, PES) = .93; r(Census, CRS) = .91 b(PES, Census) = .94; b(CRS, Census) = .91 b(Census, PES) = .93; b(Census, CRS) = .92 1950 Census r(Census, PES) = .86 b(PES, Census) = .85 b(Census, PES) = .87
Bielby, Hauser and Featherman (1977, Tables 2, 3)	Years of schooling, March 1973 CPS	1973 Occupation Changes in a Generation (OCG) and OCG re-interview (OCG-R)	mean variance CPS 12.18 8.24 OCG 11.98 11.70 OCG – R 12.12 8.58 r(CPS, OCG) = .801 r(CPS, OCG-R) = .921 r(OCG, OCGR) = .838
Ashenfelter and Krueger (1994, Tables 1, 2, 8)	Years of schooling, twins attending twin festival	Twins' report	MZ twins DZ twins variance (own report) 4.67 4.04 variance (twin report) 4.58 4.28 r(own report, twin report) .90 .91 r(Δown report, Δtwin report) .57 .74

continued on next page

Table 10, *continued*

Reference	Variables of interest	Validation source	Findings			
Kane, Rouse and Staiger (1999, Appendix Table 1)	Years of schooling, National Longitudinal Class of 1972	Transcript data	Sample proportions (self-reported data):			
			Transcript data	No College	Some College	BA+
			No College	.376	.048	.005
			Some College	.039	.271	.014
			BA+	.000	.005	.244
Miller, Mulvey and Martin (1995, Tables 1, 2)	Years of schooling, Australian Twin Register	Twins' report		MZ twins	DZ twins	
			variance (own report)	6.25	5.86	
			variance (twin report)	5.39	4.72	
			r(own report, twin report)	.88	.82	
			r(Δown report, Δtwin report)	.36	.60	
Rouse (1999, Table 1; Appendix Tables 1a,b)	Years of schooling, twins attending twin festival	Twins' report		MZ twins		
			variance (own report)	4.24		
			variance (twin report)	4.28		
			r(own report, twin report)	.92		
			r(Δown report, Δtwin report)	.62		

measures and higher variance, suggesting it is the least reliable of the three measures. A rather complicated measurement model – which allows errors to be correlated with true values for OCG and OCG-R but not CPS[81], and assumes errors in the three measures are uncorrelated – produces reliability estimates of .89, .70, and .96 for CPS, OCG, and OCG-R, respectively. The lower reliability of OCG is perhaps attributable to it being a mailback survey, while the others were telephone or personal interviews.

If we take the estimates of the first three studies in Table 10 at face value, their implication is that biases in estimating the effect of education on other variables due to errors in measuring years of schooling are not likely to be large. There are, however, two important qualifications: (i) taking these estimates at face value means assuming that the errors in the alternative reports are (at least roughly) uncorrelated, (ii) as noted in Section 2, biases due to measurement error become more important if other (well-measured) explanatory variables are correlated with years of schooling.

A relatively extreme context for illustrating the latter point are recent "twin" studies that relate wage or earnings differences between twins to differences in their schooling. In effect, this strategy for estimating returns to education adds a set of dummy variables, one for each pair of twins, to a standard wage or earnings equation. Such between-twin differencing has much the same effect as the first-differencing in panel data – most of the variation in schooling is between rather than within twin pairs, and if reporting errors are not highly correlated the reliability of differences in education within twin pairs is likely to be lower than the reliability of reports of education in general.

Ashenfelter and Krueger (1994) obtained the usual information on wages and schooling in a sample of twins, and each sample member's report of the years of schooling completed by his or her twin. This report of one's twin's schooling is highly correlated with the twin's own report ($r = .9$); assuming (as Ashenfelter and Krueger do) that errors in their own and twin reports are uncorrelated, this correlation is consistent with the reliability estimates in the earlier literature. However the correlation between twin 1's report of own schooling minus twin 2's report of own schooling and twin 2's report of 1's schooling minus twin 1's report of 2's schooling is only .57 in their sample of MZ (monozygotic, or "identical" twins) and .74 in a small sample of DZ (dizygotic, or "fraternal" twins). This suggests, for the MZ twins, that estimates of returns to schooling based on differencing wages and schooling between twins are likely to underestimate the true returns by over 40%. IV estimates, using the difference in reports of twin's schooling as an instrument for one's own reports, are consistent with this calculation[82].

[81] Bielby et al. argue that with true scores unobserved, the units of the "true" variable are arbitrary and regard the unit coefficient on the CPS measure as a normalization. Their estimates suggest a slight positive correlation between error and true value for the two OCG measures.

[82] The IV estimate reproduces this calculation if the maintained assumption that the covariance between the difference in wages and the difference in years of schooling is the same using either measure of the

The assumption that reporting errors are uncorrelated with each other is subject to challenge on a number of grounds. First, one might anticipate that the error made, for example, by twin 1 in reporting own schooling would be positively related with the error in twin 2's report of 1's schooling[83], so that errors in the own- and cross-reports of the difference in schooling would be positively related. This would lead the covariance between differences in years of schooling based on own reports and on twin reports to be greater than the variance of the true difference, and the bias due to measurement error understated by the classical model. A second possibility is that errors in one twin's report of own and twin's schooling are positively related. This would imply that twin 1's report of own schooling would be more highly correlated with his/her report of 2's schooling than with 2's report of own schooling, and the data support this conjecture. Ignoring such a correlation would lead the standard correction for the bias due to measurement error to be too large.

A solution to the second problem is to use one twin's report of the difference in schooling as an instrument for the other's report. This leads to a downward revision, as expected, in the estimated return to schooling. Behrman and Rosenzweig (1999), in contrast, find no evidence in their sample from the Minnesota Twin registry that errors in reports of own and twin's schooling are correlated, and so find estimated returns to schooling are unaffected by allowing for such a correlation.

A subsequent paper by Rouse (1999), using four waves of twin surveys rather than the first wave used by Ashenfelter and Krueger, found somewhat different substantive results[84] but quite similar conclusions as regards the importance of measurement error in the schooling variable[85].

Miller, Mulvey and Martin (1995) conducted a similar analysis using a larger sample of Australian twins. Their findings differ from the U.S. twin studies in two respects. First, the correlation between the difference in own reports and the difference in twin reports of education is substantially lower, at least for MZ twins. Second, the variance of schooling using twin reports is lower than using own reports. This suggests that the twin reports are more accurate or the errors are more mean-reverting, neither of which seem likely on a priori grounds.

difference in years of schooling. As Ashenfelter and Krueger note, this is approximately true in their data.

[83] While we lack a firm understanding of the situations which lead to errors in reporting schooling, it seems reasonable that there would be certain situations in which errors are particularly frequent, and if there is an error it is particularly likely to go in one direction. For example, if one ends one's schooling after a not-particularly-successful sophomore year of college, "true" years of schooling might be 17, with the most likely error reporting 18 instead. If twin 1 is in this situation, both twin 1 and twin 2 would be more likely to over-report schooling (by one year) than to make some other error.

[84] Unlike other twin studies, Ashenfelter and Krueger (1994) found that a first-differenced specification (not corrected for measurement error) leads to larger estimates of the returns to schooling than is obtained without fixed (twin) effects; Rouse's larger sample reaffirms the conventional wisdom in this regard.

[85] Ashenfelter and Rouse (1998) use the first three waves of the twin survey; their correlations between own and twin reports are very similar to those from Rouse's study which uses four.

Kane, Rouse and Staiger (1999) return to the “standard” framework for estimating wage equations, simple cross-sections with no (identifiable) twins. They focus instead on the assumption that the error in reporting years of schooling is unrelated to the true value. As noted above, for binary variables (e.g., has graduated from college vs. has not graduated), any error must be negatively related to the true value. The same sort of negative correlation is likely (though not inevitable) for bounded variables such as schooling.

Kane, Rouse and Staiger (1999) analyze schooling as reported by respondents in the National Longitudinal Study of the Class of 1972, virtually all of whom graduate from high school. Their focus is on reports of education beyond high school, as reported by NLS72 respondents and as recorded in transcripts of all post-secondary schools they reported attending (which were collected as part of the NLS72 study). While one might be tempted to take the latter as an indicator of “true” schooling, internal evidence suggests this is unlikely: holding constant BA receipt or non-receipt according to the transcript data, those who self-report having one earn higher wages than those who do not (and, less surprisingly, holding constant self-reported BA status, those who have a BA according to the transcript data earn higher wages than those who do not).

This provides the basis for a method-of-moments estimation strategy that does not rely on the standard IV assumption that measurement errors are uncorrelated with true values. Kane, Rouse and Staiger (1999) do, however, maintain the standard assumption that errors in reporting schooling are uncorrelated with *wages,* with each other, and (in models with covariates) with the covariates. In the simplest case, with schooling a binary variable and no covariates, there are seven unknowns: the intercept and BA-premium in the ln-wage equation, the true probability of having a BA, and four parameters of the “measurement” model (which has each measure of schooling as a linear function [with intercept] of true schooling). There are also seven observable means or sample proportions: if we define a two-by-two table for combinations of self- and transcript-reported BA status, there is one mean ln wage in each of these cells and four (but only three independent) sample proportions. This equivalence provides the basis for jointly estimating wage equations and the measurement model by GMM. Kane, Rouse, and Staiger show how this intuition can be extended to many educational categories, and to include covariates (which lead to the model being over-identified).

Substantively, they find that most differences between self-reports and transcript data – and most of the error, according to their GMM estimate – occur where one or the other of the reports claims some college, but less than a BA degree. This means that the extent to which OLS under-estimates and traditional IV overstates the return to schooling is largest as a proportion of the true value for those reporting some college. According to their estimates, OLS is less than the GMM estimate of returns to some college and a BA by about .02 (on a base of .125 and .308, respectively) while IV over-estimates each return by about the same amount [Kane, Rouse and Staiger (1999, Table 6)].

On balance, the studies in Table 10 support four general conclusions. First, evidence on the reliability of survey reports of educational attainment rely more on multiple measures, each of which is likely to contain non-negligible error, and less on direct validation evidence than is true for most of the other variables considered in this paper. Second, unless there is substantial positive correlation of the errors in these multiple measures, the bias due to errors in measuring years of schooling in traditional applications such as cross-sectional earnings functions is unlikely to be large. Third, while it is generally assumed that the errors are uncorrelated with each other and with the dependent variable (typically, ln wage or ln earnings), there is no direct evidence on this score. Most discussions in the literature treat positive correlations as the most likely alternative; if positive is more likely than negative, there is every reason to fear that positive is more likely than zero. Fourth, here as elsewhere, differencing (in this case, differences within twin pairs) greatly exacerbates the bias due to errors in measuring schooling, but the availability of reports of one's twin's schooling as well as one's own provides some leverage for undoing such bias.

7. Conclusions

Empirical research in economics has increasingly used individual- or household-level data derived from surveys. Unlike aggregate data based on surveys where one might hope that the errors would "cancel out", the move to micro data requires a continuous concern about measurement error as a likely source of bias[86]. Some variables (transfer income, wealth holdings, medical care utilization and expenditures) are sufficiently difficult to measure that such concerns would arise even in estimating simple bivariate regressions; others (union coverage, schooling, and perhaps earnings) that seem to be reported with reasonable accuracy become candidates for concern when panel data are used in ways that effectively difference out much of the true variation while increasing the noise.

The impact of measurement error on parameter estimates depends on the magnitude of the error relative to the true variation, but more generally on the joint distribution of the measurement errors and the true variables. If we are going to use data on X and y in order to study the impact of X^* on y^*, in principle we need to know the entire data-generating mechanism; that is., $f(y, X, y^*, X^*)$. Standard methods for "correcting" for measurement error such as instrumental variables procedures typically

[86] Our comments should not be taken to suggest we think aggregate data is without significant error. While response errors are presumably less important in aggregate data than they are in individual- or household-level survey data, there are certainly other important sources of error. Many aggregate series (e.g., unemployment rates) are based on survey data and, as such, are subject to sampling error. More fundamentally, much aggregate data is constructed using procedures that are likely to introduce systematic error into data series. Thus, for example the Department of Commerce's Bureau of Economic Analysis (BEA) uses procedures to construct value added [Peterson (1987)] that, outside of manufacturing and a few other industries are likely to underestimate the growth in output and thus productivity [Griliches (1994)] and to create spurious correlations between growth and productivity [Waldmann (1991)]. Any discussion of such issues is well beyond the scope of this chapter.

involve strong assumptions regarding the nature of the data-generating mechanism (i.e., that errors are classical) that are rarely discussed or defended. Short of detailed knowledge of the data-generating mechanism, the theoretical literature suggests that when the correlations between our measures and our constructs is high and when our models are simple, we can be reasonably confident regarding the robustness, in qualitative terms, of our inferences. This is the situation where standard methods for correcting for measurement error have little effect on our estimates. In contrast to this, in situations where we have reason to believe that measurement error on key variables is sufficiently large as to have qualitative effects on our estimates, serious sensitivity analysis is in order.

Validation data has provided considerable evidence on the magnitude of measurement error. Gradually, the focus has shifted from the extent of under- or overreporting (i.e., on the mean error) to the ratio of the variance of the reporting error to the variance of the true value, and more recently to consideration of whether errors are, as is so often assumed, uncorrelated with true values. Such evidence as we have suggests that errors are often negatively related to true values and, indeed, this must be so for binary variables. Fewer studies focus on the correlation between errors in measuring one variable and either measured or true values of other variables. The very limited evidence we have suggests that such correlations do not lead to appreciable or predictable biases except in contexts where variables are definitionally related (e.g., hours worked per week and earnings per hour defined as weekly earnings/weekly hours).

Despite the effort that has gone into validating various survey measures, it is striking to us how little is known about the accuracy of much of the data that is routinely collected in household surveys. To take a simple example, there is no hard evidence on how reliably hourly earnings are reported for men and women paid by the hour. Nor is there much data on the accuracy with which individuals report wealth or consumption expenditures. In other contexts, such as for health conditions, we know something about the accuracy of such reports, but virtually nothing about the impact that misreporting has on parameter estimates. Similarly, there are many studies of the accuracy of retrospective reporting of events, but few clues as to how the (often important) errors found in such studies will bias parameter estimates of event-history studies.

Increasing use of panel data has been accompanied with a heightened awareness of the tendency of such estimation to increase the importance of measurement error. The panel-data literature has benefitted from simple, intuitive results that alert analysts to situations where such errors are likely to be most harmful. Unfortunately, even the most rudimentary corrections for measurement error in such contexts depend on knowing the correlation between errors¾for an individual's wage over time, for twins' reports of their education, etc. – and there is almost no direct evidence on such correlations. Obtaining validation data sufficient to calculate such correlations requires at least two rounds of survey data and either two rounds of validation data (e.g., the PSID Validation Study) or the good fortune to be able to obtain validation of two rounds of survey data in a single step (e.g., the matched CPS–SSA data, and matches of transfer program records to SIPP data). Hopefully, in the future, it will be possible to merge administrative data to existing panel data.

As with panel data, there is good reason to fear that parameter estimates in non-linear models are likely to be more sensitive to measurement error than those in simple (linear) models. Unfortunately, the analysis of non-linear models has proceeded on a case-by-case basis, and it has not highlighted any key feature of the error distribution for validation studies to assess. Thus, analysts must often choose between less ambitious linear models for which the consequences of measurement error is better understood and more elaborate models which may well be more vulnerable to such errors. At a minimum, assessment of the relative benefits of the two approaches needs to put greater weight on this vulnerability.

One reason for remaining gaps in our knowledge about the inaccuracies of survey data is that users of the data are rarely involved in the validation studies[87]. As a result, it is natural for them to focus on the accuracy of the reports rather than the effect of inaccuracies on parameter estimates. Since different researchers are interested in different parameters, researchers conducting validation studies will never be able to satisfy all audiences. However, researchers can sometimes make their data publically available. It is interesting to note that both the CPS–SSA data by Bound and Krueger and the PSID-V data have been put to very good use by researchers outside the teams that originally developed the two data sets [e.g., Bollinger (1998), Pischke (1995), French (1998), Brownstone and Valletta (1996)]. In addition, there are clear payoffs to greater involvement of users in the design of validation studies.

While in general we believe that more effort devoted to collecting and analyzing validation data would significantly enhance the value of survey data, it is important to recognize the limitations of such initiatives. Those collecting validation data usually begin with the intention of obtaining “true” values against which the errors of survey reports can be assessed; more often than not we end up with the realization that the validation data are also imperfect. While much can still be learned from such data, particularly if one is confident the errors in the validation data are uncorrelated with those in the survey reports, this means replacing one assumption (e.g., errors are uncorrelated with true values) with another (e.g., errors in survey reports uncorrelated with errors in validation data).

Many of the validation studies reported in this chapter are based on small convenience samples (workers in a firm which cooperates by providing payroll records, households with accounts at cooperating financial institutions). The use of small samples means the reliability of the data is itself assessed with considerable sampling error. Moreover, the distribution of the variables of interest may well differ in the smaller validation sample and the large sample about which one wishes to make

[87] These comments echo somewhat similar comments often made by Griliches (e.g., 1986, 1994) that economists should become more involved in the generation of the data they use.

inferences (e.g., true wage variation will be smaller within one firm than in the economy). Even when validation data is provided for a sizeable share of a larger survey, concerns about representativeness are hard to dismiss (are those who underreport transfers less likely to cooperate in validating their responses?).

A final limitation of validation studies is that, even if the validation corresponds exactly to the "correct" answer to the survey question, it may not correspond to the "true" value of the variable in question. On the one hand, the construct we wish to test may be more subtle than questions that our surveys can ask. For example, earnings presumably depend on the interaction of years of schooling, school quality, and student effort that produce "education" or "learning"; the gap between "education" and "years of schooling" will remain no matter how successful we are in inducing individuals to accurately report their years of schooling. On the other hand, in some cases it may be the respondent's perception of a variable rather than the "true" value of the variable that motivates behavior. Thus, for example, savings behavior of smokers may depend on their own estimate of their life-expectancy, not the Surgeon General's.

It is widely recognized that survey data – and, indeed, other types of data – are often imperfect. Analyzing such data requires an understanding of their most significant shortcomings. Validation data are often imperfect, too. But they give important clues about these shortcomings – clues that would otherwise be unavailable – and suggest strategies for dealing with them. As econometricians create more complicated tools, understanding the effects of imperfect data on the performance of these tools becomes more important. Validation studies are an essential part of that enterprise.

References

Abowd, J.M., and D. Card (1987), "Intertemporal labor supply and long-term employment contracts", American Economic Review 77:50–68.

Abowd, J.M., and D. Card (1989), "On the covariance structure of earnings and hours changes", Econometrica 57:411–445.

Abowd, J.M., and A. Zellner (1985), "Estimating gross labor-force flows", Journal of Business and Economic Statistics 3:254–283.

Abrevaya, J., and J.A. Hausman (1997), "Semiparametric estimation with mismeasured dependent variables: an application to panel data on unemployment spells", Mimeo (University of Chicago, Graduate School of Business; MIT Department of Economics).

Aigner, D.J. (1973), "Regression with a binary independent variable subject to errors of observation", Journal of Econometrics 1:49–59.

Aigner, D.J., C. Hsiao, A. Kapteyn and T. Wansbeek (1984), "Latent variable models in econometrics", in Z. Griliches and M.D. Intriligator, eds., Handbook of Econometrics, Vol. II (North-Holland, Amsterdam) 1323–1393.

Akerlof, G., and J. Yellen (1985), "Unemployment through the filter of memory", Quarterly Journal of Economics 100:747–783.

Amemiya, T. (1974), "The nonlinear two-stage least-squares estimator", Journal of Econometrics 2:105–110.

Amemiya, Y. (1985), "Instrumental variable estimator for the nonlinear errors-in-variables model", Journal of Econometrics 28:273–289.

Amemiya, Y. (1990), “Two-stage instrumental variable estimators for the nonlinear errors-in-variables model”, Journal of Econometrics 44:311–332.

Andersen, E.B. (1982), “Latent structure analysis: a survey”, Scandinavian Journal of Statistics”, 9:1–12.

Anderson, K.H., and R.V. Burkhauser (1984), “The importance of the measure of health in empirical estimates of the labor supply of older men”, Economics Letters 16:375–380.

Anderson, K.H., and R.V. Burkhauser (1985), “The retirement-health nexus: a new measure of an old puzzle”, Journal of Human Resources 20:315–330.

Angrist, J., and A. Krueger (1999), “Empirical strategies in labor economics”, in: O. Ashenfelter and D. Card, eds., Handbook of Labor Economics, Vol. 3A (North-Holland, Amsterdam) 1277–1366.

Ashberg, K. (1987), “Disability as a predictor of outcome for the elderly in a department of internal medicine”, Scandinavian Journal of Social Medicine 15:26–265.

Ashenfelter, O., and A. Krueger (1994), “Estimates of the economic return to schooling from a new sample of twins”, American Economic Review 84:1157–1173.

Ashenfelter, O., and C.E. Rouse (1998), “Income, schooling, and ability: evidence from a new sample of identical twins”, Quarterly Journal of Economics 113:253–284.

Bailar, B.A. (1975), “The effects of rotation group bias on estimates from panel surveys”, Journal of the American Statistical Association 70:23–30.

Baker, M. (1997), “Growth-rate heterogeneity and the covariance structure of life-cycle earnings”, Journal of Labor Economics 15:338–375.

Baker, M., and G. Solon (1998), “Earnings dynamics and inequality among Canadian men, 1976–1992: evidence from longitudinal income tax records”, Unpublished manuscript (University of Michigan).

Bancroft, G. (1940), “Consistency of information from records and interviews”, Journal of the American Statistical Association 35:377–381.

Barron, J.M., M.C. Berger and D.A. Black (1997), On the Job Training (W.E. Upjohn Institute for Employment Research, Kalamazoo, MI).

Behrman, J.R., and M.R. Rosenzweig (1999), “Ability” biases in schooling returns and twins: a test and new estimates”, Economics of Education Review, 18:159–167.

Berger, M.C., D.A. Black and F.A. Scott (1998), “How well do we measure employer-provided health insurance coverage?” Contemporary Economic Policy 16:356–367.

Bergner, M., R. Bobbitt, S. Kressel, W. Pollard, B. Gilson and J. Morris (1976), “The sickness impact profile: conceptual formulation and methodology for the development of a health status measure”, International Journal of Health Services 6:393–415.

Bergner, M., R. Bobbitt, W. Carter and B. Gilson (1981), “The sickness impact profile: development and final revision of a health status measure”, Medical Care 19:787–805.

Bielby, W.T., R.M. Hauser and D.L. Featherman (1977), “Response errors of non-black males in models of the stratification process”, in: D.J. Aigner and A.S. Goldberger, eds., Latent Variables in Socio-Economic Models (North-Holland, Amsterdam) 227–251.

Bishop, J.H. (1974), “Biases in measurement of the productivity benefits of human capital investments”, Discussion Paper 223-74 (Institute for Research on Poverty, University of Wisconsin).

Black, D., M. Berger and F. Scott (2000), “Bounding parameter estimates with mismeasured data”, Journal of the American Statistical Association 95:739–748.

Blair, E., and S. Burton (1987), “Cognitive processes used by survey respondents to answer behavioral frequency questions”, Journal of Consumer Research 14:280–288.

Bollinger, C.R. (1996), “Bounding mean regressions when a binary regressor is mismeasured”, Journal of Econometrics 73:387–399.

Bollinger, C.R. (1998), “Measurement error in the current population survey: a nonparametric look”, Journal of Labor Economics 16:576–594.

Bollinger, C.R., and M.H. David (1997), “Measuring discrete choice with response error: food stamp participation”, Journal of the American Statistical Association 92:827–835.

Borus, M. (1966), “Response error in survey reports of earnings information”, Journal of the American Statistical Association 61:729–738.

Borus, M. (1970), “Response error and questioning technique in surveys of earnings information”, Journal of the American Statistical Association 65:566–575.

Bound, J. (1991), “Self-reported vs. Objective measures of health in retirement models”, Journal of Human Resources 26:106–138.

Bound, J., and A. Krueger (1991), “The extent of measurement error in longitudinal earnings data: do two wrongs make a right?” Journal of Labor Economics 12:345–368.

Bound, J., C. Brown, G.J. Duncan and W.L. Rodgers (1989), “Measurement error in cross-sectional and longitudinal labor market surveys: results from two validation studies”, Working Paper 2884 (National Bureau of Economic Research).

Bound, J., C. Brown, G.J. Duncan and W.L. Rodgers (1994), “Evidence on the validity of cross-sectional and longitudinal labor market data”, Journal of Labor Economics 12:345–368.

Bound, J., M. Schoenbaum and T. Waidmann (1995), “Race and education differences in disability status and labor force attachment in the health and retirement survey”, The Journal of Human Resources 30:S227–S267.

Bound, J., M. Schoenbaum, T.R. Stinebrickner and T. Waidmann (1999), “The dynamic effects of health on the labor force transitions of older workers”, Labour Economics 6:179–202.

Bowers, N., and F. Horvath (1984), “Keeping time: an analysis of errors in the measurement of unemployment duration”, Journal of Business and Economic Statistics 2:140–149.

Branden, L., and M. Pergamit (1994), “Response error in reporting starting wages”, Paper presented at the Annual Meetings of the American Association for Public Opinion Research.

Brown, C., and J. Medoff (1996), “Employer characteristics and work environment”, Annales D’Economie et de Statistique 41:275–298.

Brown, J.N., and A. Light (1992), “Interpreting panel data on job tenure”, Journal of Labor Economics 10:219–257.

Brownstone, D. (1998), “Multiple imputation methodology for missing data, non-random response and panel attrition”, in: T. Garling, T. Laitila and K. Westin, eds., Theoretical Foundations of Travel Choice Modeling (Elsevier, Amsterdam).

Brownstone, D., and R. Valletta (1996), “Modeling earnings measurement error: a multiple imputation approach”, Review of Economics and Statistics 78:705–717.

Burkhead, D., and J. Coder (1985), “Gross changes in income recipiency from the survey of income and program participation”, Proceedings of the Section on Social Statistics (American Statistical Association, Alexandria, VA) 351–356.

Burtless, G. (1987), “Occupational effects of the health and work capacity of older men”, in: G. Burtless, ed., Work, Health and Income Among the Elderly (Brookings Institution, Washington, DC).

Burton, S., and E. Blair (1991), “Task conditions, response formulation processes, and response accuracy for behavioral frequency questions in surveys”, Public Opinion Quarterly 55:50–79.

Cannell, C., and F. Fowler (1963), A Study of Reporting of Visits to Doctors in the National Health Survey (Survey Research Center, Ann Arbor, MI).

Cannell, C., G. Fisher and T. Bakker (1965), “Reporting of hospitalizations in the health interview survey”, Vital and Health Statistics, Series 2, Number 6 (Public Health Service, Washington).

Card, D. (1996), “The effect of unions on the structure of wages: a longitudinal analysis”, Econometrica 64:957–979.

Carroll, R., and L.A. Stefanski (1990), “Approximate quasi-likelihood estimation in models with surrogate predictors”, Journal of the American Statistical Association 85:652–663.

Carroll, R., J. Ruppert and L.A. Stefanski (1995), Measurement Error in Nonlinear Models (Chapman and Hall, London).

Carstensen, L., and H. Woltman (1979), “Comparing earning data from the CPS and employer’s records”, Proceedings of the Social Statistics Section (American Statistical Association, Alexandria, VA) 168–173.

Cash, W.S., and A.J. Moss (1972), “Optimum recall period for the reporting of persons injured in motor vehicle accidents", Vital and Health Statistics, Series 2, Number 5 (U.S. Public Health Service, Washington).

Chamberlain, G., and Z. Griliches (1975), "Unobservables with a variance components structure: ability, schooling and the economics success of brothers", International Economic Review 16:422–49.

Chase, D.R., and M. Harada (1984), "Response error in self-reported recreation participation", Journal of Leisure Research 16:322–329.

Chirikos, T.N., and G. Nestel (1981), "Impairment and labor market outcomes: a cross-sectional and longitudinal analysis", in: H. Parnes, ed., Work and Retirement: A Longitudinal Study of Men (MIT Press, Cambridge, MA) 93–131.

Chua, T.C., and W.Y. Fuller (1987), "A model for multinomial response error applied to labor flows", Journal of the American Statistical Association 82:46–51.

Coder, J. (1992), "Using administrative record information to evaluate the quality of the income data collected in the survey of income and program participation", Proceedings of Statistics Canada Symposium 92, Design and Analysis of Longitudinal Surveys (Statistics Canada, Ottawa) 295–306.

Cohen, S., and B. Carlson (1994), "A comparison of household and medical provider reported expenditures in the 1987 NMES", Journal of Official Statistics 10:3–29.

Cragg, J.G. (1997), "Using higher moments to estimate the simple errors-in-variables model", RAND Journal of Economics 28:S71–91.

Currie, J., and B.C. Madrian (1999), "Health, health insurance and the labor market", in: O. Ashenfelter and D. Card, eds., Handbook of Labor Economics, Vol. 3C (North-Holland, Amsterdam) 3309–3416.

Curtin, R.F., F.T. Juster and J. Morgan (1989), "Survey estimates of wealth: an assessment of quality", in: R.E. Lipsey and H.E. Tice, eds., The Measurement of Saving, Investment, and Wealth (University of Chicago Press, Chicago).

Dagenais, M.G., and D.L. Dagenais (1997), "Higher moment estimators for linear regression models with errors in the variables", Journal of Econometrics 76:193–221.

David, M. (1962), "The validity of income reported by a sample of families who received welfare assistance during 1959", Journal of The American Statistical Association 57:680–685.

Dhondt, A. (1960), "Sur une généralisation d'un theorème de R. Frisch en analyse de confluence", Cahiers du Centre d'Etudes de Recherche Operationnelle 2:37–46.

Dibbs, R., A. Hale, R. Loverock and S. Michaud (1995), "Some effects of computer assisted interviewing on the data quality of the survey of labour and income dynamics", SLID Research Paper, Series No. 95-07 (Statistics Canada, Ottawa).

Dickens, W.T., and B.A. Ross (1984), "Consistent estimation using data from more than one sample", Technical Working Paper 33 (NBER).

Dodge, R.W. (1970), "Victim recall pretest", Unpublished memorandum (U.S. Bureau of the Census, Washington, DC).

Dreher, G. (1977), "Nonrespondent characteristics and respondent accuracy in salary research", Journal of Applied Psychology 62:773–776.

Duncan, G., and D. Hill (1985), "An investigation of the extent and consequences of measurement error in labor economic survey data", Journal of Labor Economics 3:508–522.

Duncan, G., and N. Mathiowetz (1985), A Validation Study of Economic Survey Data (Survey Research Center, University of Michigan, Ann Arbor).

Ferber, R. (1966), The Reliability of Consumer Reports of Financial Assets and Debts (Bureau of Economic and Business Research, University of Illinois, Urbana, IL).

Ferber, R., J. Forsythe, H. Guthrie and E. Maynes (1969a), "Validation of a national survey of financial characteristics: savings accounts", Review of Economics and Statistics 51:436–444.

Ferber, R., J. Forsythe, H. Guthrie and E. Maynes (1969b), "Validation of consumer financial characteristics: common stock", Journal of the American Statistical Association 64:415–432.

Ferraro, K.F. (1980), "Self-ratings of health among the old and old–old", Journal of Health and Social Behavior 21:377–383.

Freeman, R.B. (1984), "Longitudinal analyses of the effects of trade unions", Journal of Labor Economics 2:1–26.

French, E. (1998), "The labor supply response to (measured but) predictable wage changes", Unpublished manuscript (Department of Economics, University of Wisconsin).

Fuller, W. (1987), Measurement Error Models (Wiley, New York).

Garber, S., and S. Klepper (1980), "Extending the classical normal errors-in-variables model", Econometrica 48:1541–1546.

Geary, R.C. (1942), "Inherent relationships between random variables", Proceedings of the Royal Irish Academy Section A 47:63–76.

Gems, B., D. Ghosh and R. Hitlin (1982), "A recall experiment: impact of time on recall of recreational fishing trips", Proceedings of the Section on Survey Research Methods (American Statistical Association, Alexandria, VA) 168–173.

Geronimus, A.T., J. Bound and L.J. Neidert (1996), "On the validity of using census geocode characteristics to proxy individual socioeconomic characteristics", Journal of the American Statistical Association 91:529–537.

Gini, C. (1921), "Sull'interpolazione di una retta quando i valori della variabile indipendente sono affetti da errori accidntali", Metron 1:63–82.

Goldberger, A.S. (1972), "Structural equation methods in the social sciences", Econometrica 40: 979–1001.

Goodman, L.A. (1974a), "The analysis of systems of qualitative variables when some of the variables are unobservable, Part I, A modified latent structure approach", American Journal of Sociology 79:1179–1259.

Goodman, L.A. (1974b), "Explanatory latent structure analysis using both identifiable and unidentifiable models", Biometrika 61:215–231.

Goodreau, K., H. Oberheu and D. Vaughan (1984), "An assessment of the quality of survey reports of income from the aid to families with dependent children (AFDC) program", Journal of Business and Economic Statistics 2:179–186.

Greenberg, D., and H. Halsey (1983), "Systematic misreporting and effects of income maintenance experiments on work effort: evidence from the Seattle–Denver experiment", Journal of Labor Economics 1:380–407.

Griliches, Z. (1974), "Errors in variables and other unobservables", Econometrica 42:971–998.

Griliches, Z. (1986), "Economic data issues", in: Z. Griliches and M.D. Intriligator, eds., Handbook of Econometrics, Vol. 3 (North-Holland, Amsterdam) 1466–1514.

Griliches, Z. (1994), "Productivity, R&D, and the data constraint", American Economic Review 84:1–23.

Griliches, Z., and J.A. Hausman (1986), "Errors in variables in panel data", Journal of Econometrics 31:93–118.

Griliches, Z., and V. Ringstad (1970), "Error in the variables bias in nonlinear contexts", Econometrica 38:368–370.

Grondin, C., and S. Michaud (1994), "Data quality of income data using computer-assisted interview: the experience of the Canadian survey of labour and income dynamics", Proceedings of the Survey Research Methods Section (American Statistical Association, Alexandria, VA) 830–835.

Groves, R.M. (1989), Survey Errors and Survey Costs (Wiley, New York).

Gustman, A.L., and T.L. Steinmeier (1999), "What people don't know about their pensions and social security: an analysis using linked data from the health and retirement study", Working Paper 7368 (NBER).

Haber, L. (1966), "Evaluating response error in the reporting of the income of the aged: benefit income", Proceedings of the Social Statistics Section (American Statistical Association, Alexandria, VA) 412–419.

Haberman, S.J. (1977), "Product models for frequency tables involving indirect observation", Annals of Statistics 5:1124–1147.

Haider, S. (2001), "Earnings instability and earnings inequality of males in the United States: 1967–1991", Journal of Labor Economics, forthcoming.

Hall, R.E., and F.S. Mishkin (1982), "Sensitivity of consumption to transitory income: estimates from panel data on households", Econometrica 50:461–481.

Halsey, H. (1978), "Validating income data: lessons from the Seattle and Denver income maintenance experiment", Proceedings of the Survey of Income and Program Participation Workshop, Survey Research Issues in Income Measurement: Field Techniques, Questionnaire Design, and Income Validation (U.S. Department of Health, Education, and Welfare, Washington, DC).

Hamermesh, D.S. (1990), "Shirking or productive schmoozing: wages and the allocation of time at work", Industrial and Labor Relations Review 43:121S–133S.

Hampel, F., E. Ronchetti, P. Rousseeuw and W. Stahel (1986), Robust Statistics (Wiley, New York).

Hardin, E., and G. Hershey (1960), "Accuracy of employee reports on changes in pay", Journal of Applied Psychology 44:269–275.

Hausman, J.A., J. Abrevaya and F. Scott-Morton (1998), "Misclassification of the dependent variable in a discrete response setting", Journal of Econometrics 87:239–269.

Heckman, J.J. (1978), "Dummy endogenous variables in a simultaneous equation system", Econometrica 46:931–959.

Heckman, J.J., and R. Robb Jr (1985), "Alternative methods for evaluating the impact of interventions", in: J.J. Heckman and B. Singer, eds., Longitudinal Analysis of Labor Market Data (Cambridge University Press, Cambridge) 156–246.

Heckman, J.J., R.J. LaLonde and J.A. Smith (1999), "The economics and econometrics of active labor market programs", in: O. Ashenfelter and D. Card, eds., Handbook of Labor Economics, Vol. 3A (North-Holland, Amsterdam) 1865–2097.

Heeringa, S.G., D.H. Hill and D.A. Howell (1995), "Unfolding brackets for reducing item nonresponse in economic surveys", Health and Retirement Study Working Paper 94-027 (Institute for Social Research, Ann Arbor, MI).

Hill, D. (1987), "Response errors around the seam: analysis of change in a panel with overlapping reference periods", Proceedings of the Section on Survey Research Methods (American Statistical Association, Alexandria, VA) 210–215.

Hoaglin, D. (1978), "Household income and income reporting error in the housing allowance demand experiment", Proceedings of the Survey of Income and Program Participation Workshop, Survey Research Issues in Income Measurement: Field Techniques, Questionnaire Design, and Income Validation (U.S. Department of Health, Education, and Welfare, Washington, DC).

Horowitz, J.L., and C.F. Manski (1995), "Identification and robustness with contaminated and corrupted data", Econometrica 63:281–302.

Horvath, F. (1982), "Forgotten unemployment: recall bias in retrospective data", Monthly Labor Review 105:40–43.

Hotz, V.J., C. Mullin and S. Sanders (1997), "Bounding causal effects using data from a contaminated natural experiment: analyzing the effects of teenage childbearing", Review of Economic Studies 64:575–603.

Hsiao, C. (1989), "Consistent estimation for some nonlinear errors-in-variables models", Journal of Econometrics 41:159–185.

Hu, T. (1971), "The validity of income and welfare information reported by a sample of welfare families", Proceedings of the Social Statistics Section (American Statistical Association, Alexandria, VA) 311–313.

Huber, P. (1981), Robust Statistics (Wiley, New York).

Hurd, M.D., and W. Rodgers (1998), "The effects of bracketing and anchoring on measurement in the health and retirement study" (Institute for Social Research, Ann Arbor, MI).

Hurd, M.D., D. McFadden, H. Chand, L. Gan, A. Merill and M. Roberts (1998), "Consumption and savings balances of the elderly: experimental evidence on survey response bias", in: D. Wise, ed., Frontiers in the Economics of Aging (University of Chicago Press, Chicago) 353–387.

Hurst, E., M.C. Luoh and F.P. Stafford (1998), "The wealth dynamics of American families, 1984–94", Brookings Papers on Economic Activity 1998:267–337.

Hwang, G.T., and L.A. Stefanski (1994), "Monotonicity of regression functions in structural measurement error models", Statistics and Probability Letters 20:113–116.

Idler, E.I., and Y. Benyamini (1997), "Self-rated health and mortality: a review of twenty-seven community studies", Journal of Health and Social Behavior 39:21–37.

Johnson, A., and M.E. Sanchez (1993), "Household and medical provider reports on medical conditions: national medical expenditure survey, 1987", Journal of Economic and Social Measurement 19: 199–223.

Juster, F.T., and J.P. Smith (1997), "Improving the quality of economic data: lessons from the HRS and AHEAD", Journal of the American Statistical Association 92:1268–1277.

Juster, F.T., J.P. Smith and F.P. Stafford (1999), "The measurement and structure of household wealth", Labour Economics 6:253–273.

Kaestner, R., T. Joyce and H. Wehbeh (1996), "The effect of maternal drug use on birth weight: measurement error in binary variables", Economic Inquiry 34:617–629.

Kane, T.J., C.E. Rouse and D. Staiger (1999), "Estimating the returns to schooling when schooling is misreported", Working Paper 7235 (NBER).

Kapteyn, A., and T. Wansbeek (1983), "Identification in the linear errors in variables model" Econometrica 51:1847–49.

Katz, S., and C. Akpom (1976), "Index of ADL", Medical Care 14:116–118.

Katz, S., A. Ford, R. Moskowitz, B. Jacobsen and M. Jaffe (1963), "Studies of illness in the aged: the index of ADL: a standardized measure of biological and psychosocial function", Journal of the American Medical Association 185:914–919.

Katz, S., T. Downs, H. Cash and R. Grotz (1970), "Progress in development of the index of ADL", Gerontologist 10:20–30.

Keating, E., D. Paterson and C. Stone (1950), "Validity of work histories obtained by interview", Journal of Applied Psychology 34:6–11.

Kish, L., and J.B. Lansing (1954), "Response errors in estimating the value of homes", Journal of the American Statistical Association 49:520–538.

Klein, B., and D. Vaughan (1980), "Validity of AFDC reporting among list frame recipients", in: J. Olson, ed., Reports from the Site Research Test (U.S. Department of Health and Human Services, Washington, DC) ch. 11.

Klepper, S., and E.E. Leamer (1984), "Consistent sets of estimates for regressions with errors in all variables", Econometrica 52:163–184.

Koopmans, T.C. (1937), Linear Regression Analysis of Economic Time Series (Netherlands Econometric Institute, de Erven F. Bohn N.V., Haarlem).

Krasker, W., and J. Pratt (1986), "Bounding the effects of proxy variable on regression coefficients", Econometrica 54:641–656.

Krueger, A.B., and L.H. Summers (1988), "Efficiency wages and the inter-industry wage structure", Econometrica 56:259–293.

Lambrinos, J. (1981), "Health: a source of bias in labor supply models", Review of Economics and Statistics 63:206–212.

Lansing, J.B., G. Ginsburg and K. Braaten (1961), An Investigation of Response Error (Bureau of Economic and Business Research, University of Illinois, Urbana IL).

LaRue, A., L. Bank, L. Jarvic and M. Hewtland (1979), "Health in old age: how physicians' rating and self-ratings compare", Journal of Gerontology 34:687–691.

Lavy, V., M. Palumbo and S. Stern (1998), "Simulation of multinomial probit probabilities and imputation", in: T. Fomby and R. Hill, eds., Advances in Econometrics (JAI Press, Greenwich CT).

Lee, L.F. (1982a), "Health and wage: a simultaneous equation model with multiple discrete indicators", International Economic Review 23:199–221.

Lee, L.F. (1982b), "Simultaneous equations models with discrete and censored variables", in: C. Manski and D. McFadden, eds., Structural Analysis of Discrete Data with Econometric Applications (MIT Press, Cambridge, MA).

Lee, L.F., and J. Sepanski (1995), "Estimation of linear and nonlinear errors-in-variables models using validation data", Journal of the American Statistical Association 90:130–140.

Levi, M.D. (1973), "Errors in the variables bias in the presence of correctly measured variables", Econometrica 41:985–986.

Levine, P. (1993), "CPS contemporaneous and retrospective unemployment compared", Monthly Labor Review 116:33–39.

Lewbel, A. (1997), "Constructing instruments for regressions with measurement error when no additional data are available, with an application to patents and R&D", Econometrica 65:1201–1213.

Little, R.J. (1992), "Regression with missing X's: a review", Journal of the American Statistical Association 87:1227–1237.

Little, R.J., and D.B. Rubin (1987), Statistical Analysis with Missing Data (Wiley, New York).

Livingston, R. (1969), "Evaluation of the reporting of public assistance income in the special census of Dane County, Wisconsin: May 15, 1968", Proceedings of the Ninth Workshop on Public Welfare Research and Statistics, 59–72.

Loeb, S., and J. Bound (1996), "The effect of measured school inputs on academic achievement: evidence from the 1920s, 1930s and 1940s birth cohorts", Review of Economics and Statistics LXXVIII:653–664.

Loftus, E.F. (1975), "Leading questions and the eyewitness report", Cognitive Psychology 7:560–572.

MaCurdy, T.E. (1982), "The use of time series processes to model the error structure of earnings in a longitudinal data analysis", Journal of Econometrics 18:83–114.

Maddala, G.S. (1983), Limited-dependent and Qualitative Variables in Econometrics (Cambridge University Press, Cambridge).

Maddox, G., and E. Douglas (1973), "Self-assessment of health: a longitudinal study of elderly subjects", Journal of Health and Social Behavior 14:87–93.

Madow, W. (1973), "Net differences in interview data on chronic conditions and information derived from medical records", Vital and Health Statistics, Series 2, No. 23 (U.S. Government Printing Office, Washington, DC).

Manning Jr, W.G., J.P. Newhouse and J.E. Ware Jr (1982), "The status of health in demand estimation, or beyond excellent, good, fair and poor", in: V.R. Fuchs, ed., Economic Aspects of Health (University of Chicago Press, Chicago).

Marquis, K.H., and J.C. Moore (1990), "Measurement errors in survey of income and program participation (SIPP) program reports", Proceedings of the Sixth Annual Research Conference (U.S. Bureau of the Census, Washington, DC) 721–745.

Mathiowetz, N. (1986), "The problem of omissions and telescoping error: new evidence from a study of unemployment", Proceedings of the Section on Survey Research Methods (American Statistical Association, Alexandria, VA).

Mathiowetz, N. (1992), "Errors in reports of occupation", Public Opinion Quarterly 56:352–355.

Mathiowetz, N., and G. Duncan (1988), "Out of work, out of mind: response errors in retrospective reports of unemployment", Journal of Business and Economic Statistics 6:221–229.

Mathiowetz, N., and T. Lair (1994), "Getting better? Change or error in the measurement of functional limitations", Journal of Economic and Social Measurement 20:237–262.

Maynes, E. (1965), "The anatomy of response errors: consumer saving." Journal of Marketing Research 2:378–387.

McCallum, B.T. (1972), "Relative asymptotic bias from errors of omission and measurement", Econometrica 40:757–758.

McHorney, C., J. Ware, J. Lu and C. Sherbourne (1994), "The MOS 36-item short-form health survey (SF-36), III Test of data quality, scaling assumptions, and reliability across diverse patient groups", Medical Care 32:40–66.

McHorney, C., M. Kosinski and J. Ware (1994), "Comparison of the costs and quality of norms collected by mail versus telephone interview: results from a national study", Medical Care 32:551–567.

Mellow, W., and H. Sider (1983), "Accuracy of response in labor market surveys: evidence and implications", Journal of Labor Economics 1:331–344.

Menon, G. (1994), "Judgements of behavioral frequencies: memory search and retrieval strategies", in: N. Schwarz and S. Sudman, eds., Autobiographical Memory and the Validity of Retrospective Reports (Springer, New York) 161–172.

Miller, H., and L. Paley (1958), "Income reported in the 1950 census and on income tax returns", An Appraisal of the 1950 Census Income Data (Princeton University Press, Princeton NJ) 179–201.

Miller, P., C. Mulvey and N. Martin (1995), "What do twin studies reveal about the economic returns to education? A comparison of Australian and U.S. findings", American Economic Review 85:586–599.

Mitchell, O. (1988), "Worker knowledge of pension provisions", Journal of Labor Economics 6(1):21–39.

Moore, J.C., and D. Kasprzyk (1984), "Month-to-month recipiency turnover in the ISDP", Proceedings of the Section on Survey Research Methods (American Statistical Association, Alexandria, VA) 210–215.

Moore, J.C., K.H. Marquis and K. Bogen (1996), "The SIPP cognitive research evaluation experiment: basic results and documentation", Unpublished report (U.S. Bureau of the Census).

Moore, J.C., L. Stinson and E. Welniak (2000), "Income measurement error in surveys: a review", Journal of Official Statistics 16(4):331–361.

Morgenstern, R., and N. Barrett (1974), "The retrospective bias in unemployment reporting by sex, race, and age", Journal of the American Statistical Association 69:355–357.

Mossey, J.M., and E. Shapiro (1982), "Self-rated health: a predictor of mortality among the elderly", American Journal of Public Health 72:800–808.

Murphy, L.R., and C.D. Cowan (1976), "Effects of bounding on telescoping in the national crime survey", Proceedings of the Social Statistics Section (American Statistical Association, Alexandria, VA) 633–638.

Muthen, B. (1983), "Latent variable structural equation modeling with categorical data", Journal of Econometrics 22:43–65.

Myers, R.J. (1982), "Why do people retire from work early?" Aging and Work 5:83–91.

Nagi, S.Z. (1969), "Congruency in medical and self-assessment of disability", Industrial Medicine 38:74–83.

National Center for Health Statistics (1961), "Health interview responses compared with medical records", Vital and Health Statistics, Series D, No. 5 (U.S. Government Printing Office, Washington, DC).

National Center for Health Statistics (1967), "Interview data on chronic conditions compared with information derived from medical records", Vital and Health Statistics, PHS Pub. No. 1000, Series 2, No. 23 (U.S. Government Printing Office, Washington, DC).

Neter, J., and J. Waksberg (1964), "A study of response errors in expenditures data from household interviews", Journal of the American Statistical Association 59:18–55.

Newey, W.K., and D. McFadden (1994), "Large sample estimation and hypothesis testing", in: R.F. Engle and D.L. McFadden, eds., Handbook of Econometrics, Vol. 4 (North-Holland, Amsterdam) 2111–2245.

Nunnally, J., and I. Bernstein (1994), Psychometric Theory (McGraw-Hill, New York).

Oberheu, H., and M. Ono (1975), "Findings from a pilot study of current and potential public assistance recipients included in the current population survey", Proceedings of the Social Statistics Section (American Statistical Association, Alexandria, VA) 576–579.

Pakes, A. (1982), "On the asymptotic bias of Wald-type estimators of a straight line when both variables are subject to error", International Economic Review 23:491–497.

Pal, M. (1980), "Consistent moment estimators of regression coefficients in the presence of errors in variables", Journal of Econometrics 14:349–364.

Parsons, D.O. (1982), "The male labor force participation decision: health, reported health, and economic incentives", Economica 49:81–91.

Patefield, W.M. (1981), “Multivariate linear relationships: maximum likelihood estimation and regression bounds”, Journal of the Royal Statistical Society B 43:342–352.

Peterson, M.O. (1987), “Gross product by industry, 1986”, Survey of Current Business 67:25–27.

Pischke, J.S. (1995), “Measurement error and earnings dynamics: some estimates from the PSID validation study”, Journal of Business and Economic Statistics 13:305–314.

Poterba, J.M., and L.S. Summers (1984), “Response variation in the CPS: caveats for the unemployment analyst”, Monthly Labor Review 107:37–42.

Poterba, J.M., and L.S. Summers (1986), “Reporting errors and labor market dynamics”, Econometrica 54:1319–1338.

Poterba, J.M., and L.S. Summers (1995), “Unemployment benefits and labor market transitions: a multinomial logit model with errors in classification”, Review of Economics and Statistics 77:207–216.

Reiersol, O. (1945), “Confluence analysis by means of instrumental sets of variables”, Arkiv for Mathematik, Astronomi och Fysik 32:1–119.

Reiersol, O. (1950), “Identifiability of a linear relation between variables which are subject to error”, Econometrica 18:375–389.

Robinson, J., and A. Bostrom (1994), “The overestimated workweek? What time-diary measures suggest”, Monthly Labor Review 117:11–23.

Rodgers, W., and B. Miller (1997), “A comparative analysis of ADL questions in surveys of older people”, The Journals of Gerontology 52B:21–36.

Rodgers, W., C. Brown and G. Duncan (1993), “Errors in survey reports of earnings, hours worked, and hourly wages”, Journal of the American Statistical Association 88:1208–1218.

Rodgers, W.L., and A.R. Herzog (1987), “Covariances of measurement errors in survey responses”, Journal of Official Statistics 3:403–418.

Rouse, C.E. (1999), “Further estimates of the economic return to schooling from a new sample of twins”, Economics of Education Review 18:149–157.

Rubin, D.B. (1987), Multiple Imputation for Nonresponse in Survey (Wiley, New York).

Schaeffer, N.C. (1994), “Errors of experience: response error in reports about child support and their implications for questionnaire design”, in: N. Schwartz and S. Sudman, eds., Autobiographical Memory and the Validity of Retrospective Reports (Springer, New York).

Shapiro, M.D. (1982), “A note on tests of the permanent income hypothesis in panel data”, Unpublished manuscript (University of Michigan).

Sickles, R.C., and P. Taubman (1986), “An analysis of the health and retirement status of the elderly”, Econometrica 54:1339–1356.

Siegal, P., and R. Hodge (1968), “A causal approach to the study of measurement error”, in: H.M. Blalock and A.B. Blalock, eds., Methodology in Social Research (McGraw-Hill, New York) 28–59.

Smith, J. (1997), “Measuring earnings levels among the poor: evidence from two samples of JTPA eligibles”, Unpublished paper (University of Western Ontario, Canada).

Solon, G. (1986), “Effects of rotation group bias on estimation of unemployment”, Journal of Business and Economic Statistics 4:105–109.

Stafford, F.P., and G. Duncan (1980), “The use of time and technology by households in the United States”, in: R. Ehrenberg, ed., Research in Labor Economics, Vol. 3 (JAI Press, Greenwich CT).

Stern, S. (1989), “Measuring the effect of disability on labor force participation”, Journal of Human Resources 24:361–395.

Stinebrickner, T.R. (1999), “Estimation of a duration model in the presence of missing data”, Review of Economics and Statistics 81:529–542.

Sudman, S., N. Bradburn and N. Schwarz (1996), Thinking about Answers: The Application of Cognitive Processes to Survey Methodology (Jossey-Bass, San Fransisco).

Theil, H. (1961), Economic Forecasts and Policy, 2nd revised edition (North-Holland, Amsterdam).

Torelli, N., and U. Trivellato (1989), “Youth unemployment duration from the Italian labor force survey”, European Economic Review 33:407–415.

Tourangeau, R. (1984), "Cognitive sciences and survey methods", in: T. Jabine, E. Loftus, M. Straf, J. Tanur and R. Tourangeau, eds., Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines (National Academy Press, Washington, DC).

Tourangeau, R., L. Rips and K. Rasinski (2000), The Psychology of Survey Response (Cambridge University Press, Cambridge).

U.S. Bureau of Labor Statistics (1983), Employer Records Analysis Survey of 1981 (Bureau of Labor Statistics, Washington).

U.S. Census Bureau (1993), "Poverty in the United States: 1992", Current Population Reports, Series P-60-185 (U.S. Census Bureau).

Vaughan, D. (1978), "Errors in reporting supplemental security income recipiency in a piolot household survey", Proceedings of the Section on Survey Research Methods (American Statistical Association, Alexandria, VA) 288–293.

Vaughan, D., and R. Yuskavage (1976), "Investigating discrepancies between social security administration and current population survey benefit data for 1972", Proceedings of the Social Statistics Section (American Statistical Association, Alexandria, VA) 824–829.

Waidmann, T., J. Bound and M. Schoenbaum (1995), "The illusion of failure: trends in the self-reported health of the U.S. elderly", Milbank Quarterly 73:253–287.

Waksberg, J., and R. Valliant (1978), Final Report on the Evaluation and Calibration of NEISS (Westat, Inc. for Consumer Products Safety Commission).

Wald, A. (1940), "The fitting of straight lines if both variables are subject to error", Annals of Mathematical Statistics 11:284–300.

Walden, D., C. Horgan and G. Cafferata (1982), "Consumer knowledge of health insurance coverage", in: C. Cannell and R. Groves, eds., Health Survey Research Methods (National Center for Health Services Research, Washington, DC).

Waldmann, R.J. (1991), "Implausible results or implausible data? Anomalies in the construction of value-added data and implications for estimates of price–cost markups", Journal of Political Economy 99:1315–1328.

Ware, J., K. Snow, M. Kosinski and B. Gandek (1993), SF-36 Health Survey: Manual and Interpretation Guide (The Health Institute, New England Medical Center, Boston).

Weinberg, C.R., D.M. Umbach and S. Greenland (1994), "When will nondifferential misclassification of an exposure preserve the direction of a trend?" American Journal of Epidemiology 140:565–571.

Weiss, D.J., R.V. Dawis, G.W. England and L.H. Lofquist (1961), Validity of Work Histories Obtained by Interview (Industrial Relations Center, University of Minnesota).

Wickens, M.R. (1972), "A note on the use of proxy variables", Econometrica 40:759–761.

Woolsey, T.D. (1953), "Results of the sick-leave memory test of October, 1952", Unpublished memorandum (Department of Health Education and Welfare).

Yaffe, R., and S. Shapiro (1979), "Medical economics survey-methods study: cost-effectiveness of alternative survey strategies", in: S. Sudman, ed., Health Survey Research Methods (National Center for Health Services Research, Washington).

Yatchew, A., and Z. Griliches (1985), "Specification error in probit models", Review of Economics and Statistics 67:134–139.

Yen, W., and H. Nelson (1996), "Testing the validity of public assistance surveys with administrative records: a validation study of welfare survey data", Paper presented at the Annual Conference of the American Association for Public Opinion Research.

This Page Intentionally Left Blank

AUTHOR INDEX

Aaronson, D. 3361
Abbring, J.H. 3417, 3419, 3432, 3435, 3439, 3441, 3449
Abowd, J.M. 3212, 3219, 3241, 3708, 3792, 3795
Abrego, L. 3694, 3695
Abrevaya, J. 3279–3282, 3726
Abrevaya, J., *see* Hausman, J.A. 3726
Adams, P. 3666
Ahn, H. 3331
Ahn, S. 3244
Ahn, S.C., *see* Schmidt, P. 3256
Ai, C. 3277
Aigner, D.J. 3725, 3732, 3733
Akerlof, G. 3304, 3354, 3795
Akpom, C., *see* Katz, S. 3822
Albert, J. 3571, 3598–3600, 3604, 3606, 3612, 3620, 3625
Albæk, K., *see* Koning, P. 3436
Albright, R. 3468, 3497
Alessie, R. 3303
Allenby, G.M. 3233, 3642
Alonso-Borrego, C. 3242, 3247, 3262, 3263
Altonji, J.G. 3219
Alvarez, J. 3264, 3265
Amaro de Matos, J.M.G. 3317, 3349, 3350, 3364
Amemiya, T. 3185, 3213, 3239, 3266, 3275, 3277, 3321, 3338–3340, 3349, 3506, 3736
Amemiya, Y. 3735–3737
An, M. 3302
Andersen, E.B. 3268, 3732
Andersen, P.K. 3386, 3398, 3399, 3423, 3430
Andersen, P.K., *see* Keiding, N. 3424, 3425
Andersen, P.K., *see* Nielsen, G.G. 3430
Anderson, E. 3304
Anderson, J.E. 3447
Anderson, K.H. 3717, 3718
Anderson, K.H., *see* Butler, J.S. 3445
Anderson, T.W. 3237, 3262, 3264
Andolfatto, D. 3670
Andreoni, J. 3303
Andrews, D.W.K. 3161, 3169, 3174, 3181, 3192, 3498
Angrist, J. 3756, 3767, 3785, 3789, 3804
Ansari, A., *see* Manchanda, P. 3642
Anti Nilsen, Ø. 3384
Antonides, G. 3384
Arellano, M. 3235, 3237, 3239, 3244, 3247, 3248, 3256, 3258, 3259, 3262, 3290
Arellano, M., *see* Alonso-Borrego, C. 3242, 3247, 3262, 3263
Arellano, M., *see* Alvarez, J. 3264, 3265
Arjas, E. 3399
Arroyo, C.R. 3426
Arthur, W.B. 3302
Ash, R. 3348
Ashberg, K. 3822
Ashenfelter, O. 3250, 3708, 3825, 3827, 3828
Athreya, K. 3168
Attanasio, O. 3693
Averintsev, M. 3308
Axelrod, R., *see* Axtell, R. 3305
Axtell, R. 3305
Axtell, R., *see* Epstein, J. 3305

Babu, G.J. 3183
Bacharach, M. 3698
Back, K. 3263
Bailar, B.A. 3792, 3799
Baker, M. 3418, 3424, 3764
Bakker, T., *see* Cannell, C. 3743, 3745, 3811, 3812
Balakrishnan, N., *see* Johnson, N.L. 3507
Baltagi, B. 3233
Bancroft, G. 3771, 3772
Banerjee, A. 3304
Bank, L., *see* LaRue, A. 3719
Barrett, N., *see* Morgenstern, R. 3793, 3800
Barron, J.M. 3756, 3769, 3785, 3788, 3807, 3809–3811
Basawa, I.V. 3193, 3194
Bassett, G. 3206
Bassett, G., *see* Koenker, R. 3206
Bauman, K. 3303

Bauwens, L. 3432
Becker, G. 3235, 3356
Becker, G.S., *see* Ghez, G.R. 3669
Behrman, J.R. 3212, 3828
Bekker, P.A. 3264
Bell, A. 3303
Bellman, R. 3482–3484, 3489, 3491
Bénabou, R. 3301
Bennett, C.H. 3628
Bentzon, M.W., *see* Andersen, P.K. 3423
Benyamini, Y., *see* Idler, E.I. 3719
Beran, R. 3161, 3162, 3166, 3183, 3185, 3210, 3211
Berger, J.O. 3599
Berger, M., *see* Black, D. 3732
Berger, M.C. 3808, 3810
Berger, M.C., *see* Barron, J.M. 3756, 3769, 3785, 3788, 3807, 3809–3811
Bergner, M. 3819, 3822, 3823
Bernardo, J.M. 3599
Bernheim, D. 3300
Bernstein, I., *see* Nunnally, J. 3823
Bertail, P. 3171
Bertrand, M. 3356
Besag, J. 3571
Best, N.G. 3596
Bhargava, A. 3235, 3239, 3249
Bickel, P.J. 3167–3169
Bielby, W.T. 3824, 3825
Bikhchandani, S. 3304
Billio, M. 3600, 3612
Binder, M. 3303, 3332, 3333, 3337
Bishop, J.H. 3824, 3825
Black, D. 3732
Black, D.A., *see* Barron, J.M. 3756, 3769, 3785, 3788, 3807, 3809–3811
Black, D.A., *see* Berger, M.C. 3808, 3810
Blair, E. 3744
Blair, E., *see* Burton, S. 3744
Blalock, H. 3301
Blanchard, O.J. 3161, 3415
Blattberg, R.C. 3604
Blau, D. 3325
Blume, L. 3300, 3304, 3305, 3313
Blundell, R. 3238, 3248, 3249, 3262, 3270
Bobbitt, R., *see* Bergner, M. 3819, 3822, 3823
Bogen, K., *see* Moore, J.C. 3771, 3775, 3777
Boizot, C. 3384
Bollinger, C.R. 3725, 3748, 3756, 3761, 3766, 3776, 3778, 3832
Bond, S. 3237
Bond, S., *see* Blundell, R. 3238, 3248, 3249, 3262
Bond, S.R., *see* Arellano, M. 3237, 3258, 3259, 3262
Bonnal, L. 3384, 3430, 3435, 3442, 3443, 3445
Bontemps, C. 3393, 3420
Booth, J.G. 3640
Borch-Johnsen, K., *see* Hougaard, P. 3424, 3425
Borgan, Ø., *see* Andersen, P.K. 3386, 3398, 3399, 3430
Borjas, G. 3356
Borjas, G.J., *see* Heckman, J.J. 3432
Borsch-Supan, A. 3472, 3498
Borus, M. 3749, 3758, 3760, 3766
Bose, A. 3189
Bostrom, A., *see* Robinson, J. 3788, 3790
Bound, J. 3711, 3717, 3735, 3748, 3752, 3755, 3758, 3760–3763, 3765, 3768, 3770, 3785, 3788, 3790, 3805, 3807
Bound, J., *see* Geronimus, A.T. 3712, 3714
Bound, J., *see* Loeb, S. 3712
Bound, J., *see* Waidmann, T. 3717
Bover, O. 3235
Bover, O., *see* Arellano, M. 3239, 3244, 3247, 3248, 3256, 3259
Bowers, N. 3794, 3801
Bowlus, A.J. 3393
Box, G.E.P. 3185, 3217
Braaten, K., *see* Lansing, J.B. 3780, 3781
Bradburn, N., *see* Sudman, S. 3743, 3744, 3766
Bradlow, E. 3642
Branden, L. 3755, 3769
Bretagnolle, J. 3352, 3424
Breusch, T.S. 3239
Brewster, K. 3337
Brock, W. 3304, 3305, 3309, 3310, 3317, 3318, 3348, 3350, 3366
Brookmeyer, R., *see* Xue, X. 3445
Brooks, S.P. 3571, 3596
Brooks-Gunn, J. 3355
Brown, B. 3187, 3188, 3302
Brown, B., *see* Steinberg, L. 3341
Brown, B.W. 3165, 3174
Brown, C. 3802, 3803, 3807, 3810
Brown, C., *see* Bound, J. 3711, 3748, 3755, 3762, 3763, 3768, 3770, 3785, 3788, 3790, 3805, 3807

Brown, C., *see* Rodgers, W. 3748, 3754, 3763, 3768, 3785, 3787, 3790
Brown, D.P., *see* Back, K. 3263
Brown, J.N. 3805
Browning, M. 3237, 3662, 3690, 3693
Browning, M., *see* Attanasio, O. 3693
Brownstone, D. 3737, 3739, 3832
Bryant, J. 3300
Buchinsky, M., *see* Andrews, D.W.K. 3174, 3181
Bühlmann, P. 3190, 3191
Bunch, D. 3558
Bunke, O. 3494
Burgess, S. 3393
Burkhauser, R.V., *see* Anderson, K.H. 3717, 3718
Burkhauser, R.V., *see* Butler, J.S. 3445
Burkhead, D. 3778
Burtless, G. 3718
Burton, S. 3744
Burton, S., *see* Blair, E. 3744
Butler, J.S. 3445
Byron, R.P. 3698

Caddy, V. 3690
Cafferata, G., *see* Walden, D. 3814, 3817
Cameron, S. 3350, 3352
Campbell, E., *see* Coleman, J. 3301
Cannell, C. 3743, 3745, 3811, 3812, 3816
Canova, F. 3233, 3693
Cao-Abad, R. 3216
Caplin, A. 3304
Card, D. 3445, 3708, 3809
Card, D., *see* Abowd, J.M. 3212, 3219, 3241, 3708
Card, D., *see* Ashenfelter, O. 3250
Carlin, B. 3608
Carlin, B., *see* Cowles, M.K. 3595, 3596
Carlin, B.P. 3571, 3573, 3599, 3600, 3603, 3614, 3616, 3633
Carlin, B.P., *see* Chib, S. 3619
Carlin, B.P., *see* Gelfand, A.E. 3598, 3619
Carlin, B.P., *see* Waller, L.A. 3600
Carlin, J.B., *see* Gelman, A. 3474, 3479
Carling, K. 3434, 3445
Carlson, B., *see* Cohen, S. 3815, 3817, 3821
Carlstein, E. 3189
Carrasco, R. 3257, 3258
Carrasco, R., *see* Arellano, M. 3290
Carroll, R. 3710–3712, 3728, 3737–3740
Carstensen, L. 3750, 3758, 3767, 3785, 3786
Carter, C. 3600, 3614, 3615
Carter, W., *see* Bergner, M. 3819, 3823
Case, A. 3302, 3337
Casella, A. 3302
Casella, G. 3571, 3589, 3593
Casella, G., *see* Robert, C.P. 3571, 3585, 3593, 3596
Cash, H., *see* Katz, S. 3818, 3822
Cash, W.S. 3743
Cavanagh, C.L. 3208, 3270
Celeux, G., *see* Robert, C.P. 3600, 3612
Chamberlain, G. XIII, 3231–3235, 3238, 3240, 3247, 3249, 3252, 3260, 3269, 3272, 3284, 3288, 3290, 3336, 3337, 3361, 3429, 3641, 3733
Chan, K.S. 3579, 3592, 3640
Chand, H., *see* Hurd, M.D. 3784
Chandra, T.K. 3176
Charlier, E. 3271, 3289
Chase, D.R. 3745
Chen, C., *see* Ai, C. 3277
Chen, M.-H. 3594, 3600, 3628
Chen, M.-H., *see* Nandram, B. 3606
Chen, R., *see* Liu, J.S. 3600
Chen, S. 3233, 3289
Chesher, A. 3214, 3215, 3437
Chia, N.-C. 3676
Chiappori, P.-A. 3283
Chib, S. 3473, 3475, 3479, 3511, 3557, 3571, 3573, 3580, 3583, 3584, 3588, 3594, 3597, 3599, 3600, 3604, 3607, 3609, 3610, 3613, 3615, 3616, 3619, 3620, 3629–3631, 3640
Chib, S., *see* Albert, J. 3571, 3599, 3600, 3604, 3606, 3612, 3620, 3625
Chib, S., *see* Carlin, B.P. 3573, 3600, 3633
Chib, S., *see* Elerian, O. 3600
Chib, S., *see* Kim, S. 3598, 3616
Chipman, H.A. 3600
Chirikos, T.N. 3718
Choi, E. 3191
Christiano, L.J. 3670
Chua, T.C. 3792, 3797
Clark, A. 3303
Clasen, D., *see* Brown, B. 3302
Clements, K.W. 3665
Clements, N., *see* Heckman, J.J. 3168
Coder, J. 3753, 3762
Coder, J., *see* Burkhead, D. 3778
Cogburn, R. 3480
Cohen, M., *see* Axtell, R. 3305
Cohen, S. 3815, 3817, 3821

Cole, H. 3303
Coleman, J. 3301, 3315, 3362
Coleman, T.S. 3426, 3434, 3445
Coleman II, W.J. 3670
Collado, M.D. 3233
Conley, T. 3354
Conlisk, J. 3303
Cook, R.J., *see* Ng, E.T.M. 3445
Cooley, T.F. 3668, 3671, 3678, 3679
Cooper, R. 3300, 3307
Cooper, S. 3301
Corcoran, M. 3355
Cosslett, S.R. 3506
Cowan, C.D., *see* Murphy, L.R. 3743
Cowles, M.K. 3595, 3596, 3606
Cowles, M.K., *see* Best, N.G. 3596
Cox, D.R. 3395, 3676
Cox, D.R., *see* Box, G.E.P. 3185, 3217
Cragg, J.G. 3729
Crane, J. 3302, 3304
Crepon, B. 3246
Cross, P. 3355
Currie, J. 3718
Curtin, R.F. 3779
Cutler, D. 3356

Dabrowska, D.M. 3404
Dagenais, D.L., *see* Dagenais, M.G. 3729
Dagenais, M.G. 3729
Daly, F., *see* Hand, D.J. 3585
Damien, P. 3641
Danise, B. 3558
Darrough, M. 3303
Dasgupta, P. 3304
Datta, S. 3168, 3191, 3193–3195
David, M. 3772
David, M.H., *see* Bollinger, C.R. 3776, 3778
David, P. 3302, 3304
David, P., *see* Dasgupta, P. 3304
Davidson, R. 3181, 3183
Davison, A.C. 3162
Dawis, R.V., *see* Weiss, D.J. 3803–3806
Dawkins, C. 3666, 3667, 3697
De Angelis, D. 3206
De Jong, D.N. 3662
de Jong, P. 3600, 3614, 3617
Deaton, A. 3233
Dellaportas, P. 3599, 3635
Dellaportas, P., *see* Brooks, S.P. 3596
DeLuca, S., *see* Rosenbaum, J. 3357
Dempster, A.P. 3639
den Haan, W.J. 3670
Denison, D.G.T. 3600
Dennis Jr, J.E. 3560
DeSarbo, W.S., *see* Young, M.R. 3642
Devereux, M.P., *see* Blundell, R. 3238
Devine, T.J. 3384, 3389
Devroye, L. 3573
DeVuyst, E.A. 3666, 3696
Dey, D., *see* Gelfand, A.E. 3573, 3600, 3627
Dey, D.K., *see* Gelfand, A.E. 3511, 3516
Dhondt, A. 3722
Diamond, P., *see* Blanchard, O.J. 3415
Dibbs, R. 3771, 3775, 3778
DiCiccio, T.J. 3573, 3628, 3631
Dickens, W.T. 3714
Diebold, F.X. 3384
Diebolt, J. 3600
Diebolt, J., *see* Robert, C.P. 3600, 3612
Diggle, P. 3622
Dixon, P.B. 3676
Dodge, R.W. 3743
Donald, S.G. 3167, 3172
Dornbusch, S., *see* Steinberg, L. 3341
Douglas, E., *see* Maddox, G. 3719
Downs, T., *see* Katz, S. 3818, 3822
Dreher, G. 3749, 3766, 3803
Ducharme, G.R., *see* Beran, R. 3162, 3166
Duesenberry, J. 3303
Duncan, G. 3305, 3318, 3748, 3751, 3760, 3763, 3765, 3768, 3785, 3787, 3790, 3796, 3799, 3801, 3805–3807, 3809, 3810
Duncan, G., *see* Brooks-Gunn, J. 3355
Duncan, G., *see* Mathiowetz, N. 3744, 3797, 3799, 3801
Duncan, G., *see* Rodgers, W. 3748, 3754, 3763, 3768, 3785, 3787, 3790
Duncan, G., *see* Solon, G. 3356
Duncan, G., *see* Stafford, F.P. 3786, 3790
Duncan, G.J., *see* Bound, J. 3711, 3748, 3755, 3762, 3763, 3768, 3770, 3785, 3788, 3790, 3805, 3807
Duncan, O., *see* Blau, D. 3325
Duneier, M. 3304
Durlauf, S. 3301, 3305, 3312, 3335
Durlauf, S., *see* Blume, L. 3304, 3313
Durlauf, S., *see* Brock, W. 3304, 3305, 3309, 3310, 3317, 3318, 3348

Earls, F., *see* Sampson, R. 3304
Eberwein, C. 3435, 3445
Efron, B. 3162, 3185, 3187

Eichenbaum, M., *see* Christiano, L.J. 3670
Eicher, S., *see* Brown, B. 3302
Eicker, F. 3215
Elbers, C. 3346, 3347, 3410, 3412
Elerian, O. 3600
Ellis, R. 3350, 3365
Elrod, T. 3496
Emirbayer, M. 3362
Engberg, J. 3392, 3445
England, G.W., *see* Weiss, D.J. 3803–3806
Engle, R. 3384, 3432
Epple, D. 3356
Epstein, J. 3305
Epstein, J., *see* Axtell, R. 3305
Erdem, T. 3489
Erkanli, A., *see* Muller, P. 3600
Escobar, M.D. 3600
Evans, W. 3361, 3362
Everett, C., *see* Freedman, D.A. 3354, 3355

Farrell, J. 3302
Featherman, D.L., *see* Bielby, W.T. 3824, 3825
Feller, W. 3409, 3411, 3413
Ferber, R. 3780–3782
Ferguson, T.S. 3506, 3507
Fernandez, R. 3356
Ferraro, K.F. 3719
Ferretti, N. 3193, 3195
Filardo, A.J. 3641
Fingleton, B. 3332
Fisher, G., *see* Cannell, C. 3743, 3745, 3811, 3812
Fisher, L., *see* Bauman, K. 3303
Fleming, T.R. 3398, 3399
Flinn, C.J. 3168, 3172, 3391, 3393, 3434, 3437, 3442, 3445
Florens, J.P. 3432, 3436
Föllmer, H. 3302, 3312
Ford, A., *see* Katz, S. 3822
Forster, J.J., *see* Dellaportas, P. 3635
Forsythe, J., *see* Ferber, R. 3780, 3782
Fougère, D., *see* Bonnal, L. 3384, 3430, 3435, 3442, 3443, 3445
Fougère, D., *see* Florens, J.P. 3432, 3436
Fowler, F., *see* Cannell, C. 3812, 3816
Francis, J., *see* Stern, R.M. 3690
Frank, R. 3303
Freedman, D.A. 3187, 3324, 3325, 3354, 3355
Freedman, D.A., *see* Bickel, P.J. 3167, 3168
Freeman, R.B. 3708, 3732, 3809
French, E. 3770, 3832
Freund, J.E. 3433
Frisvold, G.B., *see* Tsigas, M.E. 3683
Fruhwirth-Schnatter, S. 3614
Fuller, W. 3710, 3728, 3730
Fuller, W.Y., *see* Chua, T.C. 3792, 3797

Gail, M.H. 3422
Gallant, A.R. 3506
Gammerman, D. 3571, 3595, 3604
Gan, L., *see* Hurd, M.D. 3784
Gandek, B., *see* Ware, J. 3822, 3823
Garber, S. 3713, 3716
Garcia-Perez, J.I. 3392
Gawande, K. 3641
Geary, R.C. 3729
Gelfand, A.E. 3473, 3511, 3516, 3571, 3573, 3589, 3593, 3596, 3598–3600, 3619, 3627
Gelfand, A.E., *see* Carlin, B. 3608
Gelfand, A.E., *see* Mallick, B. 3599, 3600
Gelfand, A.E., *see* Wakefield, J.C. 3600, 3619
Gelfand, A.E., *see* Waller, L.A. 3600
Gelman, A. 3474, 3479–3481, 3544, 3596, 3599
Geman, D., *see* Geman, S. 3473, 3571, 3639
Geman, S. 3473, 3571, 3639
Gems, B. 3743
Genest, C. 3446
Gentle, J.E. 3573
George, E.I. 3600, 3638
George, E.I., *see* Blattberg, R.C. 3604
George, E.I., *see* Casella, G. 3571, 3589
George, E.I., *see* Chipman, H.A. 3600
Geronimus, A.T. 3712, 3714
Geweke, J. 3233, 3467, 3469, 3470, 3472, 3474, 3479–3481, 3490, 3491, 3493, 3496, 3498, 3499, 3501–3508, 3511, 3512, 3516, 3518, 3524, 3537, 3538, 3540–3542, 3558, 3561, 3571, 3572, 3580, 3583, 3594, 3596, 3638
Geyer, C. 3593
Geyer, C.J. 3480, 3598
Geyer, C.J., *see* Chan, K.S. 3579
Ghez, G.R. 3669
Ghosh, D., *see* Gems, B. 3743
Ghosh, J.K., *see* Chandra, T.K. 3176
Ghysels, E. 3616
Gilks, W.R. 3571
Gill, R.D. 3167

Gill, R.D., *see* Andersen, P.K. 3386, 3398, 3430
Gill, R.D., *see* Nielsen, G.G. 3430
Gilson, B., *see* Bergner, M. 3819, 3822, 3823
Gini, C. 3721
Ginsburg, G., *see* Lansing, J.B. 3780, 3781
Giot, P., *see* Bauwens, L. 3432
Glaeser, E. 3304, 3305, 3312, 3353
Glaeser, E., *see* Cutler, D. 3356
Godsill, S.J. 3635
Goering, J. 3358
Goldberger, A.S. 3506, 3733
Gönül, F. 3430
Goodman, L.A. 3354, 3732
Goodreau, K. 3774, 3778
Goodwin, J., *see* Emirbayer, M. 3362
Goolsbee, A. 3302
Gordon, R., *see* Corcoran, M. 3355
Gordon, S.F., *see* Filardo, A.J. 3641
Gottfredson, M. 3304
Gottschalk, P., *see* Engberg, J. 3445
Götze, F. 3193
Götze, F., *see* Bickel, P.J. 3169
Granger, C.W.J. 3240
Granovetter, M. 3301–3303
Green, E., *see* Besag, J. 3571
Green, P.E. 3573, 3600, 3633, 3635, 3636
Green, P.J., *see* Richardson, S. 3635
Greenberg, D. 3750
Greenberg, E., *see* Chib, S. 3473, 3475, 3479, 3557, 3571, 3580, 3583, 3584, 3588, 3597, 3600, 3604, 3610, 3615, 3620, 3631, 3640
Greene, W.H. 3506
Greenland, S., *see* Robins, J.M. 3412
Greenland, S., *see* Weinberg, C.R. 3722, 3727
Greenwood, J. 3671, 3678, 3679
Gregory, A.W. 3655, 3662
Griffith, R., *see* Blundell, R. 3270
Griliches, Z. 3212, 3237, 3714, 3727, 3733, 3830, 3832
Griliches, Z., *see* Chamberlain, G. 3733
Griliches, Z., *see* Hausman, J.A. 3270
Griliches, Z., *see* Yatchew, A. 3728
Gritz, R.M. 3435, 3445
Grondin, C. 3771, 3775, 3777, 3779, 3780, 3783
Grossman, M., *see* Becker, G. 3235
Grotz, R., *see* Katz, S. 3818, 3822
Groves, R.M. 3746
Guo, G. 3426, 3430, 3446
Gupta, S., *see* Manchanda, P. 3642
Gustman, A.L. 3810
Guthrie, H., *see* Ferber, R. 3780, 3782

Haber, L. 3772
Haberman, S.J. 3732
Hahn, J. 3187, 3206, 3260, 3264, 3424, 3430
Haider, S. 3764
Hajivassiliou, V. 3233, 3288, 3470, 3472, 3498, 3502
Hajivassiliou, V., *see* Borsch-Supan, A. 3472, 3498
Hale, A., *see* Dibbs, R. 3771, 3775, 3778
Hall, B., *see* Hausman, J.A. 3270
Hall, P. 3161, 3162, 3165, 3167, 3168, 3171, 3176, 3177, 3180–3183, 3185, 3187–3192, 3199–3204, 3210, 3222
Hall, P., *see* Choi, E. 3191
Hall, P., *see* De Angelis, D. 3206
Hall, R.E. 3212, 3241, 3708
Halsey, H. 3771, 3773, 3777, 3778
Halsey, H., *see* Greenberg, D. 3750
Ham, J.C. 3426, 3435, 3437, 3445
Ham, J.C., *see* Eberwein, C. 3435, 3445
Hamermesh, D.S. 3787, 3790
Hamilton, B., *see* Chib, S. 3600
Hamilton, J.D. 3612
Hammersly, J.M. 3473
Hampel, F. 3724
Han, A.K. 3279, 3281, 3282
Hand, D.J. 3585
Handscomb, D.C., *see* Hammersly, J.M. 3473
Hannan, E.J. 3480
Hansen, G.D., *see* Cooley, T.F. 3671, 3678, 3679
Hansen, L.P. 3192, 3256, 3263, 3332
Hansen, L.P., *see* Browning, M. 3662, 3690, 3693
Hansen, T., *see* Scarf, H.E. 3657
Harada, M., *see* Chase, D.R. 3745
Harberger, A.C. 3657, 3666, 3673, 3685
Hardin, E. 3749, 3766
Härdle, W. 3161, 3202, 3203, 3216, 3280, 3285
Harrington, D.P., *see* Fleming, T.R. 3398, 3399
Harris, R., *see* Cox, D.R. 3676
Harrison, G.W. 3666, 3696
Harvald, B., *see* Anderson, J.E. 3447
Harvald, B., *see* Hougaard, P. 3427, 3447
Harvey, A.C., *see* Ghysels, E. 3616
Hastings, W.K. 3473, 3475, 3571, 3583, 3587

Hauser, R. 3318
Hauser, R.M., *see* Bielby, W.T. 3824, 3825
Hausman, J.A. 3239, 3270, 3726
Hausman, J.A., *see* Abrevaya, J. 3726
Hausman, J.A., *see* Griliches, Z. 3237, 3714
Haveman, R. 3302
Hayashi, F. 3238, 3256
Heaton, J., *see* Hansen, L.P. 3263
Heckman, J.J. 3168, 3272, 3276, 3283, 3284, 3301, 3329, 3337, 3338, 3341–3344, 3346, 3347, 3352, 3357, 3384, 3398, 3405, 3406, 3408, 3410, 3411, 3414, 3418, 3419, 3424, 3432, 3434, 3435, 3439–3442, 3445, 3449, 3492, 3504, 3519, 3521, 3530, 3732, 3733
Heckman, J.J., *see* Browning, M. 3662, 3690, 3693
Heckman, J.J., *see* Cameron, S. 3350, 3352
Heckman, J.J., *see* Chen, S. 3233
Heckman, J.J., *see* Flinn, C.J. 3168, 3172, 3391, 3434, 3437, 3442, 3445
Heeringa, S.G. 3784
Henderson, J.V. 3360
Hercowitz, Z., *see* Greenwood, J. 3678, 3679
Hershey, G., *see* Hardin, E. 3749, 3766
Hertel, T.W. 3681, 3683
Herzog, A.R., *see* Rodgers, W.L. 3762, 3780, 3782
Hewtland, M., *see* LaRue, A. 3719
Hidalgo, J., *see* Baltagi, B. 3233
Higdon, D., *see* Besag, J. 3571
Higgs, P., *see* Adams, P. 3666
Hildenbrand, W., *see* Härdle, W. 3161
Hill, D. 3745, 3778
Hill, D., *see* Duncan, G. 3748, 3751, 3760, 3763, 3765, 3768, 3785, 3787, 3790, 3796, 3799, 3801, 3805, 3806, 3809, 3810
Hill, D.H., *see* Heeringa, S.G. 3784
Hill, R. 3695
Hillier, G.H. 3212, 3263
Hills, S., *see* Gelfand, A.E. 3599
Hills, S.E. 3598
Hinkley, D.V., *see* Davison, A.C. 3162
Hipp, C., *see* Götze, F. 3193
Hirano, K., *see* Chamberlain, G. 3233, 3641
Hirschi, T., *see* Gottfredson, M. 3304
Hirshleifer, D., *see* Bikhchandani, S. 3304
Hitlin, R., *see* Gems, B. 3743
Hoaglin, D. 3771, 3773, 3778
Hobert, J.P., *see* Booth, J.G. 3640
Hobson, J., *see* Coleman, J. 3301
Hodge, R., *see* Siegal, P. 3725, 3824, 3825
Hodrick, R.J. 3670
Hoeting, J.A., *see* Raftery, A.E. 3356, 3637
Hoffer, T., *see* Coleman, J. 3362
Hoffman, S. 3335
Hoffman, S., *see* Plotnick, R. 3361
Holm, A., *see* Van den Berg, G.J. 3435
Holm, N.V., *see* Anderson, J.E. 3447
Holm, N.V., *see* Hougaard, P. 3427, 3447
Holtz-Eakin, D. 3237, 3240, 3241
Hommes, C., *see* Brock, W. 3305
Honoré, B.E. 3273, 3274, 3276, 3281, 3282, 3284–3287, 3290, 3336, 3337, 3414, 3428, 3441, 3442, 3448
Honoré, B.E., *see* Heckman, J.J. 3439–3441
Hoover, K.D. 3662, 3663
Horgan, C., *see* Walden, D. 3814, 3817
Horowitz, J.L. 3162, 3174, 3184, 3206–3210, 3212, 3214–3221, 3233, 3270, 3271, 3396, 3410, 3424, 3425, 3506, 3724
Horowitz, J.L., *see* Hall, P. 3187, 3188, 3190–3192
Horvath, F. 3793, 3800
Horvath, F., *see* Bowers, N. 3794, 3801
Horvath, W.J. 3384
Hotz, V.J. 3490, 3724
Hotz, V.J., *see* Heckman, J.J. 3435, 3442, 3445
Hougaard, P. 3411, 3413, 3420, 3424, 3425, 3427, 3431, 3443, 3447
Houser, D. 3494
Houser, D., *see* Geweke, J. 3493, 3538, 3540, 3541
Howell, D.A., *see* Heeringa, S.G. 3784
Hsiao, C. 3233, 3736
Hsiao, C., *see* Aigner, D.J. 3733
Hsiao, C., *see* Anderson, T.W. 3237, 3264
Hsiao, C., *see* Li, Q. 3233
Hu, L. 3287
Hu, L., *see* Honoré, B.E. 3287
Hu, T. 3772
Huber, P. 3494, 3724
Huber-Carol, C., *see* Bretagnolle, J. 3352, 3424
Huet, S., *see* Härdle, W. 3203
Huff, K.M., *see* Young, L.M. 3683
Huffman, G., *see* Greenwood, J. 3671
Huggett, M. 3678, 3680, 3683
Hurd, M.D. 3784
Hurst, E. 3784
Hwang, G.T. 3727

Iachine, I.A., *see* Yashin, A.I. 3434, 3447
Ichimura, H. 3506
Ichimura, H., *see* Heckman, J.J. 3357
Idler, E.I. 3719
Imbens, G. 3263
IMSL 3560
Ingram, B.F., *see* De Jong, D.N. 3662
Inoue, T., *see* Hayashi, F. 3238
Insua, D.R., *see* Muller, P. 3600
Ioannides, Y. 3307, 3356

Jacobsen, B., *see* Katz, S. 3822
Jacobson, T., *see* Carling, K. 3434, 3445
Jacquier, E. 3616
Jacubson, G. 3289
Jaffe, M., *see* Katz, S. 3822
Jain, D.C., *see* Vilcassim, N.J. 3384
Janas, D. 3206
Jarvic, L., *see* LaRue, A. 3719
Jeffreys, H. 3573
Jeliazkov, I., *see* Chib, S. 3573, 3631
Jen, L., *see* Allenby, G.M. 3642
Jeong, J. 3162
Jerison, M., *see* Härdle, W. 3161
Jewitt, I., *see* Chesher, A. 3215
Jing, B.-Y., *see* Hall, P. 3171, 3190
Johansen, L. 3670
John, A., *see* Cooper, R. 3300, 3307
Johnson, A. 3819, 3821
Johnson, N.L. 3507, 3520, 3553, 3563
Johnson, P., *see* Imbens, G. 3263
Jolivet, E., *see* Härdle, W. 3203
Jones, A. 3303
Jones, C.S. 3641
Jones, L.E. 3671
Jones, R.C., *see* Harrison, G.W. 3666, 3696
Jones, S. 3300
Jorgenson, D.W. 3665
Jovanovic, B. 3300, 3320, 3389
Joyce, T., *see* Kaestner, R. 3711
Judd, C., *see* Krosnick, J. 3303
Judd, K.L. 3491, 3670
Junker, B.W., *see* Patz, R.J. 3600
Juster, F.T. 3779, 3780, 3784
Juster, F.T., *see* Curtin, R.F. 3779

Kac, M. 3365
Kadane, J., *see* Tierney, L. 3572
Kaestner, R. 3711
Kalaba, R., *see* Bellman, R. 3484, 3489, 3491
Kalbfleisch, J.D. 3386, 3398, 3426, 3429
Kane, T.J. 3732, 3733, 3826, 3829
Kao, C. 3233
Kapteyn, A. 3303, 3729
Kapteyn, A., *see* Aigner, D.J. 3733
Kapteyn, A., *see* Alessie, R. 3303
Karim, M.R., *see* Zeger, S.L. 3479, 3600, 3619
Kasprzyk, D., *see* Moore, J.C. 3778
Kass, R.E., *see* DiCiccio, T.J. 3573, 3628, 3631
Katz, L. 3359
Katz, L., *see* Case, A. 3302
Katz, M. 3302
Katz, S. 3818, 3822
Keane, M. 3233, 3236, 3288, 3467, 3472, 3483, 3484, 3487–3491, 3495, 3498, 3499, 3502, 3504, 3505, 3538
Keane, M., *see* Elrod, T. 3496
Keane, M., *see* Erdem, T. 3489
Keane, M., *see* Geweke, J. 3233, 3467, 3490, 3491, 3493, 3496, 3498, 3499, 3501–3506, 3518, 3537, 3538, 3540, 3541, 3558
Keating, E. 3749, 3766
Kehoe, T.J. 3675
Keiding, N. 3413, 3424, 3425, 3431
Keiding, N., *see* Andersen, P.K. 3386, 3398, 3430
Kelly, M. 3302
Kendall, M.G. 3446
Kennan, J.F. 3162, 3214, 3384
Kerkhofs, M., *see* Lindeboom, M. 3426
Kiefer, J. 3551
Kiefer, N.M. 3386, 3641
Kiefer, N.M., *see* An, M. 3302
Kiefer, N.M., *see* Bowlus, A.J. 3393
Kiefer, N.M., *see* Devine, T.J. 3384, 3389
Kilgore, S., *see* Coleman, J. 3362
Kim, C.-J. 3641
Kim, J. 3208, 3270
Kim, S. 3598, 3616
Kimbell, L.J., *see* Harrison, G.W. 3666, 3696
King, G. 3354, 3642
King, R.G. 3670, 3671
Kirman, A. 3301, 3307
Kish, L. 3780, 3781
Kitamura, Y. 3188
Kitcher, P. 3304
Kiviet, J.F. 3262
Klaassen, C.A.J. 3424
Klebanov, P., *see* Brooks-Gunn, J. 3355
Klein, B. 3774

Klein, J.P., *see* Andersen, P.K. 3423
Klein, J.P., *see* Keiding, N. 3424, 3425
Klein, J.P., *see* Li, Y. 3425
Klein, J.P., *see* Pelz, C.J. 3385
Klein, J.P., *see* Wang, S.T. 3426
Klein, R.W. 3506
Klein, S., *see* Freedman, D.A. 3354, 3355
Klenow, P., *see* Goolsbee, A. 3302
Klepper, S. 3722, 3723
Klepper, S., *see* Garber, S. 3713, 3716
Kling, J., *see* Katz, L. 3359
Kloek, T. 3469, 3572
Kocherlakota, N.R. 3670
Koenker, R. 3206
Koenker, R., *see* Bassett, G. 3206
Kohn, R., *see* Carter, C. 3600, 3614, 3615
Kohn, R., *see* Shively, T.S. 3600
Kohn, R., *see* Smith, M. 3638
Kollman, K. 3303
Kong, A., *see* Liu, J.S. 3592, 3593, 3597
Koning, P. 3436
Koop, G. 3641
Koopmans, T.C. 3687, 3722
Kortram, R.A. 3410, 3424
Kosinski, M., *see* McHorney, C. 3820, 3823
Kosinski, M., *see* Ware, J. 3822, 3823
Kotkin, B., *see* Bellman, R. 3484, 3489, 3491
Kotz, S., *see* Johnson, N.L. 3507, 3520, 3553, 3563
Kramarz, F., *see* Crepon, B. 3246
Kranton, R., *see* Akerlof, G. 3304
Krasker, W. 3723
Kreiss, J.-P. 3190
Kremer, M. 3356
Kressel, S., *see* Bergner, M. 3822
Krosnick, J. 3303
Krueger, A., *see* Angrist, J. 3756, 3767, 3785, 3789, 3804
Krueger, A., *see* Ashenfelter, O. 3708, 3825, 3827, 3828
Krueger, A., *see* Bound, J. 3748, 3752, 3758, 3760–3763, 3765
Krueger, A.B. 3708, 3804
Krugman, P. 3302
Krusell, P., *see* Greenwood, J. 3678, 3679
Krystal, A.D., *see* West, M. 3641
Kuhn, B., *see* Tsigas, M.E. 3683
Kuhn, T. 3304
Kunitomo, N. 3264
Kunitomo, N., *see* Anderson, T.W. 3262
Künsch, H.R. 3189
Künsch, H.R., *see* Götze, F. 3193
Kuo, L. 3638
Kydland, F.E. 3658, 3667
Kyriazidou, E. 3271, 3275, 3276, 3278, 3281, 3282, 3287, 3337
Kyriazidou, E., *see* Honoré, B.E. 3274, 3276, 3282, 3284, 3285, 3287, 3336, 3337

Labov, W. 3301
Ladd, H. 3359
Lahiri, S. 3193
Lahiri, S.N. 3190–3193
Lair, T., *see* Mathiowetz, N. 3822
Laird, N.M. 3619
Laird, N.M., *see* Dempster, A.P. 3639
LaLonde, R.J., *see* Eberwein, C. 3435, 3445
LaLonde, R.J., *see* Ham, J.C. 3435, 3437, 3445
LaLonde, R.J., *see* Heckman, J.J. 3449, 3732
Lam, K., *see* So, M.K.P. 3641
Lambrinos, J. 3718
Lancaster, T. 3214, 3270, 3337, 3386, 3389, 3394, 3396, 3401, 3405, 3406, 3415, 3418, 3422, 3423, 3427, 3430, 3435, 3439, 3544, 3641
Lancaster, T., *see* Chesher, A. 3437
Lansing, J.B. 3780, 3781
Lansing, J.B., *see* Kish, L. 3780, 3781
Laren, D., *see* Corcoran, M. 3355
LaRue, A. 3719
Laub, J., *see* Sampson, R. 3304, 3325
Lavy, V. 3736
Lawless, J., *see* Qin, J. 3263
Lawrance, E. 3693
Leahy, J., *see* Caplin, A. 3304
Leamer, E.E. 3356, 3599
Leamer, E.E., *see* Klepper, S. 3722, 3723
Ledolter, J., *see* Chan, K.S. 3640
Lee, H. 3682, 3684
Lee, L.F. 3733, 3738
Lee, L.-F. 3493, 3497–3499, 3504, 3505
Lee, M.-J. 3289
Legros, P. 3356
Lehmann, E.L. 3161
Lenk, P.J. 3600
Lenstra, A.J. 3410
Lenstra, A.J., *see* Klaassen, C.A.J. 3424
Lenstra, A.J., *see* Kortram, R.A. 3410, 3424
Leone, R.P., *see* Allenby, G.M. 3642
Lerman, S. 3468, 3497
Lerman, S., *see* Albright, R. 3468, 3497

Levi, M.D. 3713
Levine, P. 3798, 3800, 3801
Lewbel, A. 3506, 3729
Lewbel, A., *see* Honoré, B.E. 3290
Lewis, S.M., *see* Raftery, A.E. 3596
Li, H. 3195
Li, K. 3641
Li, Q. 3233
Li, Q., *see* Baltagi, B. 3233
Li, W.K., *see* So, M.K.P. 3641
Li, Y. 3425
Liang, K.-Y., *see* Diggle, P. 3622
Liebman, J., *see* Katz, L. 3359
Light, A., *see* Brown, J.N. 3805
Lillard, L.A. 3241, 3384, 3426, 3434, 3435, 3437, 3445, 3518
Lindbeck, A. 3304
Lindeboom, M. 3426, 3443
Lindeboom, M., *see* Van den Berg, G.J. 3384, 3433, 3434, 3445
Lindstrom, D.P. 3384
Linton, O., *see* Härdle, W. 3280, 3285
Little, R.J. 3737, 3739
Liu, C. 3640
Liu, J.S. 3592, 3593, 3597, 3600
Liu, R.Y. 3185, 3216
Livingston, R. 3772, 3777
Loeb, S. 3712
Lofquist, L.H., *see* Weiss, D.J. 3803–3806
Loftus, E.F. 3746
Louis, T.A. 3640
Louis, T.A., *see* Anderson, J.E. 3447
Louis, T.A., *see* Carlin, B.P. 3571
Loury, G. 3302
Loverock, R., *see* Dibbs, R. 3771, 3775, 3778
Lu, J., *see* McHorney, C. 3819, 3823
Lucas Jr, R.E. 3662, 3671
Ludwig, J., *see* Ladd, H. 3359
Lukacs, E. 3367
Lunn, A.D., *see* Hand, D.J. 3585
Luoh, M.C., *see* Hurst, E. 3784
Luttmer, E., *see* Bertrand, M. 3356

MacKay, J., *see* Genest, C. 3446
MacKinnon, J.G. 3216
MacKinnon, J.G., *see* Davidson, R. 3181, 3183
MaCurdy, T.E. 3241, 3708
MaCurdy, T.E., *see* Amemiya, T. 3239
MaCurdy, T.E., *see* Heckman, J.J. 3272
Maddala, G.S. 3506, 3728
Maddala, G.S., *see* Jeong, J. 3162
Maddala, G.S., *see* Li, H. 3195
Maddox, G. 3719
Madigan, A.D., *see* Raftery, A.E. 3356, 3637
Madow, W. 3818, 3821
Madrian, B.C., *see* Currie, J. 3718
Magnac, T. 3284
Mailath, G., *see* Cole, H. 3303
Mallick, B. 3599, 3600
Mallick, B., *see* Kuo, L. 3638
Mallick, B.K., *see* Denison, D.G.T. 3600
Mallik, A.K., *see* Basawa, I.V. 3193, 3194
Mammen, E. 3162, 3167, 3216
Mammen, E., *see* Härdle, W. 3216
Manchanda, P. 3642
Manning Jr, W.G. 3719
Manski, C.F. 3172, 3205, 3208, 3270, 3271, 3280–3282, 3286, 3305, 3318, 3319, 3321–3323, 3325, 3330, 3331, 3333, 3336, 3341, 3343–3346, 3490, 3506
Manski, C.F., *see* Albright, R. 3468, 3497
Manski, C.F., *see* Cross, P. 3355
Manski, C.F., *see* Horowitz, J.L. 3724
Manski, C.F., *see* Lerman, S. 3468, 3497
Mansur, A. 3662, 3672, 3690
Manton, K.G., *see* Vaupel, J.W. 3394
Manuelli, R.E., *see* Jones, L.E. 3671
Marcet, A. 3491
Marcet, A., *see* Canova, F. 3233
Marcet, A., *see* den Haan, W.J. 3670
Marinari, E. 3598
Markatou, M., *see* Horowitz, J.L. 3233
Marquis, K.H. 3771, 3774, 3777, 3778
Marquis, K.H., *see* Moore, J.C. 3771, 3775, 3777
Marron, J.S., *see* Härdle, W. 3216
Martin, A.D., *see* Quinn, K.M. 3642
Martin, G. 3641
Martin, N., *see* Miller, P. 3826, 3828
Martin, W., *see* Yang, Y. 3683
Mathiowetz, N. 3744, 3797, 3799, 3801–3804, 3822
Mathiowetz, N., *see* Duncan, G. 3805, 3807
May, S., *see* Brown, B. 3187, 3188
Maynes, E. 3780, 3781
Maynes, E., *see* Ferber, R. 3780, 3782
McCall, B.P. 3414
McCallum, B.T. 3713
McCausland, W., *see* Geweke, J. 3518
McConway, K.J., *see* Hand, D.J. 3585

McCormick, W.P., *see* Basawa, I.V. 3193, 3194
McCormick, W.P., *see* Datta, S. 3191
McCulloch, R.E. 3599, 3600, 3612
McCulloch, R.E., *see* Chipman, H.A. 3600
McCulloch, R.E., *see* George, E.I. 3600, 3638
McCullogh, C.E., *see* Newman, J.L. 3426
McDougall, R. 3683
McFadden, D. 3321, 3349, 3366, 3497, 3498, 3501, 3502
McFadden, D., *see* Hajivassiliou, V. 3233, 3472, 3498
McFadden, D., *see* Hurd, M.D. 3784
McFadden, D., *see* Newey, W.K. 3368, 3728
McGrattan, E.R. 3670
McHorney, C. 3819, 3820, 3823
McKitrick, R.R. 3665
McManus, D. 3326, 3327
McPartland, J., *see* Coleman, J. 3301
Meaden, P., *see* Popkin, S. 3357
Medoff, J., *see* Brown, C. 3802, 3803, 3807, 3810
Meghir, C. 3445
Meghir, C., *see* Bond, S. 3237
Mehra, R. 3670
Melenberg, B., *see* Charlier, E. 3271, 3289
Melino, A. 3410
Melino, A., *see* Baker, M. 3418, 3424
Mellow, W. 3751, 3767, 3769, 3785, 3786, 3790, 3802–3804, 3806, 3809
Meng, X.L., *see* Gelman, A. 3599
Meng, X.-L. 3573, 3628
Mengersen, K.L. 3479, 3583
Mengersen, K.L., *see* Besag, J. 3571
Menon, G. 3744
Merill, A., *see* Hurd, M.D. 3784
Metropolis, N. 3473, 3475, 3571, 3582, 3583, 3639
Meyer, B.D. 3386
Meyn, S.P. 3577, 3596
Michaud, S., *see* Dibbs, R. 3771, 3775, 3778
Michaud, S., *see* Grondin, C. 3771, 3775, 3777, 3779, 3780, 3783
Mieszkowski, R., *see* Henderson, J.V. 3360
Migon, H.S., *see* Gammerman, D. 3604
Milgrom, P. 3300
Milhaud, X., *see* Bunke, O. 3494
Miller, B., *see* Rodgers, W. 3822
Miller, H. 3749, 3758
Miller, J., *see* Kollman, K. 3303
Miller, M.I., *see* Snyder, D.L. 3431
Miller, P. 3826, 3828
Miller, R.A., *see* Hotz, V.J. 3490
Miller, S., *see* Rosenbaum, J. 3357
Min, C., *see* Zellner, A. 3481, 3596
Mira, A. 3641
Mishkin, F.S., *see* Hall, R.E. 3212, 3241, 3708
Mitchell, O. 3810, 3811
Miyao, T. 3362
Mizon, G.E., *see* Breusch, T.S. 3239
Moeschberger, M.L., *see* Li, Y. 3425
Moeschberger, M.L., *see* Wang, S.T. 3426
Moffitt, R. 3233, 3305, 3318, 3324
Molotch, H., *see* Duneier, M. 3304
Monahan, J.C., *see* Andrews, D.W.K. 3192
Monfort, A., *see* Billio, M. 3600, 3612
Montgomery, A.L. 3642
Montgomery, J. 3301
Mood, A., *see* Coleman, J. 3301
Moon, H.R., *see* Phillips, P.C.B. 3233
Moore, J.C. 3771, 3775–3778, 3783
Moore, J.C., *see* Marquis, K.H. 3771, 3774, 3777, 3778
Morgan, J., *see* Curtin, R.F. 3779
Morgenstern, R. 3793, 3800
Morimune, K. 3264
Morris, J., *see* Bergner, M. 3822
Morris, S. 3300
Mortensen, D.T. 3389, 3390, 3392, 3393, 3403, 3420
Morwitz, V.G., *see* Young, M.R. 3642
Moscarini, G. 3422
Moskowitz, R., *see* Katz, S. 3822
Moss, A.J., *see* Cash, W.S. 3743
Mossey, J.M. 3719
Mroz, T. 3445
Mullainathan, S., *see* Bertrand, M. 3356
Muller, P. 3600
Mullin, C., *see* Hotz, V.J. 3724
Mulvey, C., *see* Miller, P. 3826, 3828
Munshi, K. 3337
Murphy, K., *see* Becker, G. 3235
Murphy, L.R. 3743
Murray, J. 3366
Muthen, B. 3733
Myaux, J., *see* Munshi, K. 3337
Myers, R.J. 3717
Myglegaard, P., *see* Hougaard, P. 3424, 3425

Nagi, S.Z. 3719
Nandram, B. 3606

Nardari, F., *see* Chib, S. 3616
Narendranathan, W. 3391
National Center for Health Statistics 3743, 3818, 3821
Neal, D. 3362
Neal, D., *see* Heckman, J.J. 3301
Nechyba, T. 3304
Neidert, L.J., *see* Geronimus, A.T. 3712, 3714
Nelson, C.R. 3212
Nelson, C.R., *see* Kim, C.-J. 3641
Nelson, H., *see* Yen, W. 3771, 3775, 3777
Nestel, G., *see* Chirikos, T.N. 3718
Neter, J. 3743
Neumann, G.R. 3386
Neumann, G.R., *see* Bowlus, A.J. 3393
Neumann, G.R., *see* Kennan, J.F. 3162, 3214
Newey, W.K. 3192, 3288, 3289, 3368, 3728
Newey, W.K., *see* Brown, B. 3187, 3188
Newey, W.K., *see* Holtz-Eakin, D. 3237, 3240, 3241
Newhouse, J.P., *see* Manning Jr, W.G. 3719
Newman, A., *see* Legros, P. 3356
Newman, J.L. 3426
Newton, M.A. 3627
Neyman, J. 3267
Ng, E.T.M. 3445
Nguyen, T.T. 3676
Nickell, S.J. 3264, 3395
Nickell, S.J., *see* Narendranathan, W. 3391
Nielsen, G.G. 3430
Nijman, T. 3289
Ntzoufras, I., *see* Dellaportas, P. 3635
Nummelin, E. 3577
Nunnally, J. 3823
Nyberg, S., *see* Lindbeck, A. 3304
Nychka, D.W., *see* Gallant, A.R. 3506

Oakes, D. 3446, 3447
Oates, W., *see* Evans, W. 3361
Oberheu, H. 3771, 3773, 3777
Oberheu, H., *see* Goodreau, K. 3774, 3778
Oddou, C., *see* Kirman, A. 3307
O'Hagan, A. 3599
Omori, Y. 3435
Ono, M., *see* Oberheu, H. 3771, 3773, 3777
Oomes, N. 3304
Orme, C. 3162, 3214
Ostland, M., *see* Freedman, D.A. 3354
Ostrowski, E., *see* Hand, D.J. 3585
Oswald, A., *see* Clark, A. 3303

Paap, R. 3641
Paarsch, H.J., *see* Donald, S.G. 3167, 3172
Pagan, A.R. 3666, 3696
Page, M., *see* Solon, G. 3356
Page, S. 3309
Page, S., *see* Kollman, K. 3303
Pakes, A. 3335, 3384, 3497, 3729
Pal, M. 3729
Paley, L., *see* Miller, H. 3749, 3758
Palumbo, M., *see* Lavy, V. 3736
Panis, C.W.A., *see* Lillard, L.A. 3426, 3434, 3435, 3445
Paparoditis, E. 3191
Parente, S.L. 3680, 3683
Parisi, G., *see* Marinari, E. 3598
Parmenter, B.R., *see* Dixon, P.B. 3676
Parsons, D.O. 3717, 3718
Pastor, L. 3641
Patefield, W.M. 3722
Paterson, D., *see* Keating, E. 3749, 3766
Patz, R.J. 3600
Pelz, C.J. 3385
Pepper, J., *see* Manski, C.F. 3341, 3344–3346
Percy, D.F. 3604
Perez, J.F., *see* Amaro de Matos, J.M.G. 3317, 3349, 3350, 3364
Pergamit, M., *see* Branden, L. 3755, 3769
Perroni, C. 3666
Perroni, C., *see* Nguyen, T.T. 3676
Pesaran, M.H. 3233, 3685
Pesaran, M.H., *see* Binder, M. 3303, 3332, 3333, 3337
Peskun, P.H. 3473
Petersen, J.H. 3434
Peterson, M.O. 3830
Pfanzagl, J. 3494
Phillips, P.C.B. 3212, 3233, 3262
Piantadosi, S., *see* Gail, M.H. 3422
Piggott, J. 3676, 3681, 3683, 3684, 3687, 3690, 3691, 3694
Pischke, J.S. 3764, 3832
Pissarides, C.A. 3390
Pissarides, C.A., *see* Mortensen, D.T. 3420
Pitt, M.K. 3600, 3615
Plosser, C.I., *see* King, R.G. 3670
Plotnick, R. 3361
Plotnick, R., *see* Hoffman, S. 3335
Poirier, D.J. 3599
Politis, D.N. 3170, 3171, 3189
Politis, D.N., *see* Bertail, P. 3171
Politis, D.N., *see* Paparoditis, E. 3191

Pollak, R., *see* Darrough, M. 3303
Pollard, D., *see* Kim, J. 3208, 3270
Pollard, D., *see* Pakes, A. 3497
Pollard, W., *see* Bergner, M. 3822
Polson, N.G. 3596
Polson, N.G., *see* Carlin, B.P. 3599, 3600, 3603, 3614, 3616
Polson, N.G., *see* Jacquier, E. 3616
Popkin, S. 3357
Popkin, S., *see* Rosenbaum, J. 3357
Porter, J. 3280
Postlewaite, A. 3303
Postlewaite, A., *see* Cole, H. 3303
Poterba, J.M. 3732, 3792, 3794, 3796, 3799, 3801
Potter, S.M., *see* Koop, G. 3641
Powell, J.L. 3207, 3231, 3233, 3273, 3276, 3281, 3282, 3506
Powell, J.L., *see* Ahn, H. 3331
Powell, J.L., *see* Amemiya, T. 3185
Powell, J.L., *see* Honoré, B.E. 3281, 3282
Prado, R., *see* West, M. 3641
Pratt, J., *see* Krasker, W. 3723
Prebisch, R. 3688
Preckel, P.V., *see* DeVuyst, E.A. 3666, 3696
Prentice, R.L., *see* Kalbfleisch, J.D. 3386, 3398, 3426, 3429
Prescott, E.C. 3664
Prescott, E.C., *see* Cooley, T.F. 3668, 3678, 3679
Prescott, E.C., *see* Hodrick, R.J. 3670
Prescott, E.C., *see* Kydland, F.E. 3658, 3667
Prescott, E.C., *see* Mehra, R. 3670
Prescott, E.C., *see* Parente, S.L. 3680, 3683
Propp, J.G. 3641

Qin, J. 3263
Quah, D., *see* Blanchard, O.J. 3161
Quah, D., *see* Durlauf, S. 3335
Quinn, K.M. 3642

Racine-Poon, A., *see* Gelfand, A.E. 3599
Racine-Poon, A., *see* Wakefield, J.C. 3600, 3619
Raftery, A.E. 3356, 3596, 3637
Raftery, A.E., *see* DiCiccio, T.J. 3573, 3628, 3631
Raftery, A.E., *see* Newton, M.A. 3627
Rajarshi, M.B. 3191
Rao, C.R. 3161
Rasch, G. 3268
Rasinski, K., *see* Tourangeau, R. 3746
Rauch, J., *see* Casella, A. 3302
Raudenbusch, S., *see* Duncan, G. 3305, 3318
Raudenbusch, S., *see* Sampson, R. 3304
Rea, S.A., *see* Ham, J.C. 3426
Reagan, P., *see* Weinberg, B. 3360
Rebelo, S.T., *see* King, R.G. 3670, 3671
Rebelo, S.T., *see* Stokey, N.L. 3671
Rees, D., *see* Zax, J. 3355
Reeves, J.H., *see* Basawa, I.V. 3193, 3194
Reiersol, O. 3722, 3728
Renault, E., *see* Ghysels, E. 3616
Richards, T., *see* Trussell, J. 3418, 3424
Richardson, S. 3635
Richardson, S., *see* Gilks, W.R. 3571
Ridder, G. 3396, 3398, 3405, 3410–3412, 3420, 3422–3424, 3426, 3429, 3431
Ridder, G., *see* Elbers, C. 3346, 3347, 3410, 3412
Ridder, G., *see* Koning, P. 3436
Ridder, G., *see* Kortram, R.A. 3410, 3424
Ridder, G., *see* Van den Berg, G.J. 3393, 3433, 3434, 3445
Ringstad, V., *see* Griliches, Z. 3727
Ripley, B. 3573, 3580
Ripley, R.D. 3473
Rips, L., *see* Tourangeau, R. 3746
Ritter, C. 3596, 3630
Rivkin, S. 3355
Robb, R., *see* Heckman, J.J. 3342
Robb Jr, R., *see* Heckman, J.J. 3732
Robert, C.P. 3571, 3585, 3593, 3595, 3596, 3600, 3612
Robert, C.P., *see* Billio, M. 3600, 3612
Robert, C.P., *see* Casella, G. 3593
Robert, C.P., *see* Diebolt, J. 3600
Roberts, B.M. 3666
Roberts, D.J., *see* Milgrom, P. 3300
Roberts, G.O. 3476, 3478, 3592, 3596, 3597, 3641
Roberts, G.O., *see* Brooks, S.P. 3596
Roberts, G.O., *see* Smith, A.F.M. 3571
Roberts, M., *see* Freedman, D.A. 3354
Roberts, M., *see* Hurd, M.D. 3784
Robin, J.M., *see* Boizot, C. 3384
Robin, J.M., *see* Bontemps, C. 3393, 3420
Robins, J.M. 3412, 3449
Robinson, J. 3788, 3790
Robinson, P.M. 3289
Rodgers, W. 3748, 3754, 3763, 3768, 3785, 3787, 3790, 3822

Rodgers, W., *see* Hurd, M.D. 3784
Rodgers, W.L. 3762, 3780, 3782
Rodgers, W.L., *see* Bound, J. 3711, 3748, 3755, 3762, 3763, 3768, 3770, 3785, 3788, 3790, 3805, 3807
Rodríguez, G., *see* Guo, G. 3426, 3430, 3446
Rogerson, R., *see* Fernandez, R. 3356
Rogerson, R., *see* McGrattan, E.R. 3670
Rogerson, R., *see* Rupert, P. 3670
Roland-Holst, D.W., *see* Lee, H. 3682, 3684
Romano, J.P., *see* Bertail, P. 3171
Romano, J.P., *see* Politis, D.N. 3170, 3171, 3189
Romer, D. 3304
Romer, T., *see* Epple, D. 3356
Romo, J., *see* Ferretti, N. 3193, 3195
Ronchetti, E., *see* Hampel, F. 3724
Rosen, H., *see* Holtz-Eakin, D. 3237, 3240, 3241
Rosen, O., *see* King, G. 3642
Rosenbaum, J. 3357
Rosenbaum, J., *see* Popkin, S. 3357
Rosenbluth, A.W., *see* Metropolis, N. 3473, 3475, 3571, 3582, 3583, 3639
Rosenbluth, M.N., *see* Metropolis, N. 3473, 3475, 3571, 3582, 3583, 3639
Rosenthal, J.S. 3596
Rosenthal, J.S., *see* Roberts, G.O. 3641
Rosenzweig, M.R., *see* Behrman, J.R. 3212, 3828
Rosholm, M. 3434
Ross, B.A., *see* Dickens, W.T. 3714
Rosser, J.B. 3301
Rossi, P.E., *see* Allenby, G.M. 3233
Rossi, P.E., *see* Jacquier, E. 3616
Rossi, P.E., *see* Jones, L.E. 3671
Rossi, P.E., *see* Montgomery, A.L. 3642
Rothenberg, T. 3327
Rouse, C.E. 3359, 3826, 3828
Rouse, C.E., *see* Ashenfelter, O. 3828
Rouse, C.E., *see* Kane, T.J. 3732, 3733, 3826, 3829
Rousseeuw, P., *see* Hampel, F. 3724
Roy, A.D. 3492
Rubin, D.B. 3598, 3737
Rubin, D.B., *see* Dempster, A.P. 3639
Rubin, D.B., *see* Gelman, A. 3474, 3479–3481, 3596, 3599
Rubin, D.B., *see* Little, R.J. 3737
Rubin, D.B., *see* Liu, C. 3640
Rudebusch, G.D., *see* Diebold, F.X. 3384
Runkle, D.E. 3161, 3236
Runkle, D.E., *see* Geweke, J. 3496, 3498, 3499, 3501–3505, 3537, 3558
Runkle, D.E., *see* Keane, M. 3236, 3490
Rupert, P. 3670
Ruppert, J., *see* Carroll, R. 3710–3712, 3728, 3738–3740
Russell, J., *see* Engle, R. 3384, 3432
Rust, J. 3389, 3467, 3483–3485, 3489
Rutherford, T.F., *see* Perroni, C. 3666
Ruud, P.A., *see* Hajivassiliou, V. 3288, 3470, 3472, 3498
Ryu, K. 3431, 3437

Sacerdote, B., *see* Glaeser, E. 3304, 3305, 3312, 3353
Sacks, P., *see* Freedman, D.A. 3354, 3355
Sah, R. 3304
Sahu, S.K., *see* Gelfand, A.E. 3598, 3619
Sahu, S.K., *see* Roberts, G.O. 3597
Salanie, B., *see* Chiappori, P.-A. 3283
Saloner, G., *see* Farrell, J. 3302
Sampson, R. 3304, 3325
Sanchez, M.E., *see* Johnson, A. 3819, 3821
Sanders, S., *see* Hotz, V.J. 3724
Santos, M.S. 3484, 3487
Sargan, J.D. 3263
Sargan, J.D., *see* Bhargava, A. 3235, 3239, 3249
Sargent, T., *see* Hansen, L.P. 3332
Sastry, N. 3426, 3434
Sattinger, M. 3356
Sauvageau, Y., *see* Henderson, J.V. 3360
Sawa, T., *see* Anderson, T.W. 3262
Scarf, H.E. 3657
Schaeffer, N.C. 3744
Schankerman, M., *see* Pakes, A. 3384
Scheinkman, J., *see* Glaeser, E. 3304, 3305, 3312, 3353
Schelling, T. 3302
Schiantarelli, F., *see* Anti Nilsen, Ø. 3384
Schiantarelli, F., *see* Blundell, R. 3238
Schmidt, P. 3256
Schmidt, P., *see* Ahn, S. 3244
Schmidt, P., *see* Breusch, T.S. 3239
Schnabel, R.B., *see* Dennis Jr, J.E. 3560
Schoenbaum, M., *see* Bound, J. 3735
Schoenbaum, M., *see* Waidmann, T. 3717
Scholz, J.K., *see* Andreoni, J. 3303
Schumacher, B., *see* Stern, R.M. 3690
Schwab, R., *see* Evans, W. 3361, 3362

Schwarz, N., *see* Sudman, S. 3743, 3744, 3766
Scott, E.L., *see* Neyman, J. 3267
Scott, F., *see* Black, D. 3732
Scott, F.A., *see* Berger, M.C. 3808, 3810
Scott-Morton, F., *see* Hausman, J.A. 3726
Sealand, N., *see* Brooks-Gunn, J. 3355
Sedlacek, G., *see* Heckman, J.J. 3492
Segal, L.M., *see* Altonji, J.G. 3219
Sepanski, J., *see* Lee, L.F. 3738
Sérandon, A., *see* Bonnal, L. 3384, 3430, 3435, 3442, 3443, 3445
Shannon, J.H., *see* Pagan, A.R. 3666, 3696
Shao, Q.-M., *see* Chen, M.-H. 3600, 3628
Shao, U. 3162
Shapiro, C., *see* Katz, M. 3302
Shapiro, E., *see* Mossey, J.M. 3719
Shapiro, M.D. 3708
Shapiro, S., *see* Yaffe, R. 3813, 3816
Shephard, N. 3615, 3616
Shephard, N., *see* Chib, S. 3616
Shephard, N., *see* de Jong, P. 3600, 3614, 3617
Shephard, N., *see* Elerian, O. 3600
Shephard, N., *see* Kim, S. 3598, 3616
Shephard, N., *see* Pitt, M.K. 3600, 3615
Sherbourne, C., *see* McHorney, C. 3819, 3823
Shively, T.S. 3600
Shoven, J.B. 3657, 3672
Sickles, R.C. 3718
Sider, H., *see* Mellow, W. 3751, 3767, 3769, 3785, 3786, 3790, 3802–3804, 3806, 3809
Sieg, H., *see* Epple, D. 3356
Siegal, P. 3725, 3824, 3825
Silverman, B.W. 3196
Sims, C.A. 3239, 3324, 3662
Sims, C.A., *see* Hayashi, F. 3256
Singer, B., *see* Heckman, J.J. 3337, 3338, 3346, 3347, 3352, 3405, 3406, 3410, 3411, 3418, 3419, 3424
Singer, H.W. 3688
Singh, K. 3183
Singh, K., *see* Babu, G.J. 3183
Singh, K., *see* Liu, R.Y. 3185
Singleton, K., *see* Hansen, L.P. 3192
Slavin, R. 3360
Slesnick, D.T., *see* Jorgenson, D.W. 3665
Smith, A. 3642
Smith, A.F.M. 3571
Smith, A.F.M., *see* Bernardo, J.M. 3599
Smith, A.F.M., *see* Carlin, B. 3608
Smith, A.F.M., *see* Dellaportas, P. 3599
Smith, A.F.M., *see* Denison, D.G.T. 3600
Smith, A.F.M., *see* Gelfand, A.E. 3473, 3571, 3589, 3593, 3596, 3598, 3599
Smith, A.F.M., *see* Hills, S.E. 3598
Smith, A.F.M., *see* Roberts, G.O. 3476, 3478, 3592
Smith, A.F.M., *see* Wakefield, J.C. 3600, 3619
Smith, G.W., *see* Gregory, A.W. 3655, 3662
Smith, J. 3758
Smith, J., *see* Heckman, J.J. 3168, 3357
Smith, J.A., *see* Heckman, J.J. 3449, 3732
Smith, J.P., *see* Juster, F.T. 3779, 3780, 3784
Smith, M. 3638
Smith, R., *see* Pesaran, M.H. 3233
Smythe, J., *see* Freedman, D.A. 3354, 3355
Snow, K., *see* Ware, J. 3822, 3823
Snyder, D.L. 3431
So, M.K.P. 3641
Solomon, P.J. 3425
Solon, G. 3356, 3404, 3792
Solon, G., *see* Baker, M. 3764
Solon, G., *see* Corcoran, M. 3355
Soong, R., *see* Granovetter, M. 3302, 3303
Sørensen, T.I.A., *see* Nielsen, G.G. 3430
Spady, R., *see* Imbens, G. 3263
Spady, R.H., *see* Klein, R.W. 3506
Spiegelhalter, D.J., *see* Gilks, W.R. 3571
Spitzer, F. 3308
Spitzer, J.J. 3217
Srinivasan, K., *see* Gönül, F. 3430
St. Hilaire, F. 3676
Stafford, F.P. 3786, 3790
Stafford, F.P., *see* Hurst, E. 3784
Stafford, F.P., *see* Juster, F.T. 3779
Stahel, W., *see* Hampel, F. 3724
Staiger, D., *see* Kane, T.J. 3732, 3733, 3826, 3829
Stallard, E., *see* Vaupel, J.W. 3394
Stambaugh, R.F., *see* Pastor, L. 3641
Startz, R., *see* Nelson, C.R. 3212
Steel, M.F.J., *see* Kiefer, N.M. 3641
Stefanski, L.A., *see* Carroll, R. 3710–3712, 3728, 3737–3740
Stefanski, L.A., *see* Hwang, G.T. 3727
Steinberg, L. 3341
Steinmeier, T.L., *see* Gustman, A.L. 3810
Stengos, T., *see* Li, Q. 3233
Stephens, D.A. 3599, 3608
Stern, H.S., *see* Gelman, A. 3474, 3479, 3599
Stern, R.M. 3690

Stern, S. 3489, 3735
Stern, S., *see* Lavy, V. 3736
Stevens, J., *see* Geweke, J. 3518
Stinebrickner, T.R. 3736
Stinebrickner, T.R., *see* Bound, J. 3735
Stinson, L., *see* Moore, J.C. 3776, 3783
Stock, J.H., *see* Powell, J.L. 3506
Stoffer, D.S., *see* Carlin, B.P. 3599, 3600, 3614, 3616
Stoker, T.M., *see* Powell, J.L. 3506
Stokey, N.L. 3671
Stone, C., *see* Keating, E. 3749, 3766
Stone, R. 3698
Streufert, P. 3301
Stute, W. 3161
Sucoff, C. 3304, 3337
Sudman, S. 3743, 3744, 3766
Sueyoshi, G.T., *see* Melino, A. 3410
Sullivan, D., *see* Card, D. 3445
Summers, L.H., *see* Krueger, A.B. 3708, 3804
Summers, L.S., *see* Poterba, J.M. 3732, 3792, 3794, 3796, 3799, 3801
Sutton, J., *see* Dixon, P.B. 3676
Swanepoel, J.W.H. 3169

Taber, C.R., *see* Heckman, J.J. 3398, 3408, 3414, 3424
Tanner, M.A. 3473, 3571, 3589, 3590, 3592, 3595
Tanner, M.A., *see* King, G. 3642
Tanner, M.A., *see* Ritter, C. 3596, 3630
Tanner, M.A., *see* Wei, G.C.G. 3607, 3640
Tapia, R.A. 3469
Taubman, P., *see* Behrman, J.R. 3212
Taubman, P., *see* Sickles, R.C. 3718
Taylor, L.W. 3162, 3214
Taylor, R.L., *see* Basawa, I.V. 3193, 3194
Taylor, S.J. 3616
Taylor, W.E., *see* Hausman, J.A. 3239
Teller, A.H., *see* Metropolis, N. 3473, 3475, 3571, 3582, 3583, 3639
Teller, E., *see* Metropolis, N. 3473, 3475, 3571, 3582, 3583, 3639
Theil, H. 3716
Thomas, J.M. 3437
Thompson, E.A., *see* Geyer, C.J. 3598
Thompson, J.R., *see* Tapia, R.A. 3469
Thompson, T.S., *see* Manski, C.F. 3208
Tibshirani, R.J., *see* Efron, B. 3162
Tierney, L. 3476–3480, 3510, 3571, 3572, 3577–3580, 3583, 3598
Tierney, L., *see* Mira, A. 3641
Todd, P., *see* Heckman, J.J. 3357
Topa, G. 3353
Topa, G., *see* Conley, T. 3354
Torelli, N. 3797, 3801
Tourangeau, R. 3743, 3746
Traub, J.F. 3489
Trivellato, U., *see* Torelli, N. 3797, 3801
Trognon, A., *see* Crepon, B. 3246
Trussell, J. 3418, 3424
Tsay, R., *see* McCulloch, R.E. 3599, 3600, 3612
Tsigas, M.E. 3683
Tsionas, E.G. 3600
Tu, D., *see* Shao, U. 3162
Tunalı, I., *see* Ridder, G. 3398, 3426, 3429, 3431
Tweedie, R.L., *see* Mengersen, K.L. 3479, 3583
Tweedie, R.L., *see* Meyn, S.P. 3577, 3596
Tweedie, R.L., *see* Roberts, G.O. 3596
Tyers, R., *see* McDougall, R. 3683

U.S. Bureau of Labor Statistics 3784
U.S. Census Bureau 3765, 3779
Udry, C., *see* Honoré, B.E. 3282
Umbach, D.M., *see* Weinberg, C.R. 3722, 3727
Upchurch, D., *see* Sucoff, C. 3304, 3337
Uzawa, H., *see* Koopmans, T.C. 3687

Valletta, R., *see* Brownstone, D. 3739, 3832
Valliant, R., *see* Waksberg, J. 3745
van de Geer, S., *see* Kapteyn, A. 3303
van de Stadt, H., *see* Kapteyn, A. 3303
Van den Berg, G.J. 3384, 3391–3393, 3403, 3405, 3412, 3413, 3420, 3425, 3433–3435, 3437, 3444, 3445
Van den Berg, G.J., *see* Abbring, J.H. 3417, 3419, 3435, 3439, 3441, 3449
Van den Berg, G.J., *see* Bontemps, C. 3393, 3420
Van den Berg, G.J., *see* Koning, P. 3436
Van den Berg, G.J., *see* Lindeboom, M. 3443
Van den Berg, G.J., *see* Ridder, G. 3420, 3422
Van der Klaauw, B., *see* Van den Berg, G.J. 3435
Van der Klaauw, W. 3389
van Dijk, H.K., *see* Kloek, T. 3469, 3572
van Dijk, H.K., *see* Paap, R. 3641
Van Ours, J.C., *see* Abbring, J.H. 3435

Van Ours, J.C., *see* Van den Berg, G.J. 3425, 3435
Van Rooij, A.C.M., *see* Kortram, R.A. 3410, 3424
Van Rooij, A.C.M., *see* Lenstra, A.J. 3410
van Soest, A., *see* Charlier, E. 3271, 3289
van Zwet, W.R., *see* Bickel, P.J. 3169
Vaughan, D. 3773, 3774
Vaughan, D., *see* Goodreau, K. 3774, 3778
Vaughan, D., *see* Klein, B. 3774
Vaupel, J.W. 3394
Verbeek, M., *see* Nijman, T. 3289
Verbrugge, R. 3304
Verdinelli, I. 3573, 3600, 3628
Vigo-Aguiar, J., *see* Santos, M.S. 3484, 3487
Vijverberg, W.P.M. 3498
Vilcassim, N.J. 3384
Vincent, D.P., *see* Dixon, P.B. 3676
Viner, J. 3672
Vines, S.K., *see* Best, N.G. 3596
Vinod, H.D. 3162
Vinod, H.D., *see* Harrison, G.W. 3666, 3696
Visser, M. 3431
Visser, M., *see* Boizot, C. 3384
Vytlacil, E., *see* Chen, S. 3233

Wahba, S., *see* Chia, N.-C. 3676
Waidmann, T. 3717
Waidmann, T., *see* Bound, J. 3735
Wakefield, J., *see* Damien, P. 3641
Wakefield, J.C. 3600, 3619
Waksberg, J. 3745
Waksberg, J., *see* Neter, J. 3743
Wald, A. 3729
Walden, D. 3814, 3817
Waldmann, R.J. 3830
Wales, T., *see* Darrough, M. 3303
Walker, J.R., *see* Heckman, J.J. 3384, 3435, 3442, 3445
Walker, S., *see* Damien, P. 3641
Waller, L.A. 3600
Wallis, K. XIII, 3331, 3332
Wang, S.T. 3426
Wansbeek, T., *see* Aigner, D.J. 3733
Wansbeek, T., *see* Kapteyn, A. 3303, 3729
Ware, J. 3822, 3823
Ware, J., *see* McHorney, C. 3819, 3820, 3823
Ware Jr, J.E., *see* Manning Jr, W.G. 3719
Ware, J.H., *see* Laird, N.M. 3619
Wasilkowski, G.W., *see* Traub, J.F. 3489
Wasserman, L., *see* DiCiccio, T.J. 3573, 3628, 3631
Wasserman, L., *see* Verdinelli, I. 3573, 3600, 3628
Watson, M.W. 3662, 3670
Weber, G., *see* Attanasio, O. 3693
Weber, S., *see* Kirman, A. 3307
Wehbeh, H., *see* Kaestner, R. 3711
Wei, G.C.G. 3607, 3640
Weibull, J., *see* Lindbeck, A. 3304
Weinberg, B. 3360
Weinberg, C.R. 3722, 3727
Weinfeld, F., *see* Coleman, J. 3301
Weir, D., *see* Mroz, T. 3445
Weiss, D.J. 3803–3806
Welch, I., *see* Bikhchandani, S. 3304
Welniak, E., *see* Moore, J.C. 3776, 3783
West, K.D. 3161
West, K.D., *see* Newey, W.K. 3192
West, M. 3641
West, M., *see* Escobar, M.D. 3600
West, M., *see* Muller, P. 3600
Whalley, J. 3664, 3676, 3689
Whalley, J., *see* Abrego, L. 3694, 3695
Whalley, J., *see* Chia, N.-C. 3676
Whalley, J., *see* Kehoe, T.J. 3675
Whalley, J., *see* Mansur, A. 3662, 3672, 3690
Whalley, J., *see* Piggott, J. 3676, 3681, 3683, 3684, 3687, 3690, 3691, 3694
Whalley, J., *see* Shoven, J.B. 3657, 3672
Whalley, J., *see* St. Hilaire, F. 3676
White, H. 3161, 3213–3215
White, H., *see* MacKinnon, J.G. 3216
Whitehouse, E., *see* Meghir, C. 3445
Whiteman, C.H., *see* De Jong, D.N. 3662
Whitford, A.B., *see* Quinn, K.M. 3642
Wickens, M.R. 3713
Wieand, S., *see* Gail, M.H. 3422
Wiese, A.M. 3667
Wigle, R.M. 3666, 3696
Wigle, R.M., *see* Harrison, G.W. 3666, 3696
Wigle, R.M., *see* Nguyen, T.T. 3676
Wigle, R.M., *see* Whalley, J. 3676
Wilcoxen, P.J., *see* Jorgenson, D.W. 3665
Willis, R.J., *see* Lillard, L.A. 3241, 3518
Wilson, D.B., *see* Propp, J.G. 3641
Windmeijer, F., *see* Blundell, R. 3270
Winkelmann, R., *see* Chib, S. 3597, 3640
Witte, J. 3359
Wolf, D., *see* Engberg, J. 3445
Wolf, M., *see* Politis, D.N. 3170, 3171

Wolfe, B., *see* Haveman, R. 3302
Wolfowitz, J., *see* Kiefer, J. 3551
Wolpin, K.I. 3391, 3392
Wolpin, K.I., *see* Keane, M. 3467, 3483, 3484, 3487–3489, 3491, 3538
Woltman, H., *see* Carstensen, L. 3750, 3758, 3767, 3785, 3786
Wong, W.H., *see* Liu, J.S. 3592, 3593, 3597
Wong, W.H., *see* Meng, X.-L. 3573, 3628
Wong, W.H., *see* Tanner, M.A. 3473, 3571, 3589, 3590, 3592
Wood, S., *see* Shively, T.S. 3600
Wooldridge, J.M. 3252, 3289
Woolsey, T.D. 3743
Woutersen, T.M. 3431, 3448
Wozniakowski, H., *see* Traub, J.F. 3489
Wright, R., *see* McGrattan, E.R. 3670
Wright, R., *see* Rupert, P. 3670
Wu, C.F.J. 3216, 3640
Wu, Y.N., *see* Liu, C. 3640
Wyhowski, D., *see* Schmidt, P. 3256

Xia, H., *see* Waller, L.A. 3600
Xue, X. 3445

Yaffe, R. 3813, 3816
Yamaguchi, K. 3386, 3429
Yanagishima, K., *see* Yang, Y. 3683
Yang, Y. 3683
Yankow, J., *see* Weinberg, B. 3360
Yaron, A., *see* Hansen, L.P. 3263
Yashin, A.I. 3434, 3447
Yatchew, A. 3728
Yellen, J., *see* Akerlof, G. 3304, 3795
Yen, W. 3771, 3775, 3777
Yinger, J. 3357
Yoon, B.J. 3391, 3402
York, R., *see* Coleman, J. 3301
Young, G.A., *see* De Angelis, D. 3206
Young, H.P. 3300
Young, L.M. 3683
Young, M.R. 3642
Yuskavage, R., *see* Vaughan, D. 3773

Zabel, J.E. 3289
Zahl, P.H. 3434
Zaslavsky, A.M., *see* Bradlow, E. 3642
Zax, J. 3355
Zeger, S.L. 3479, 3600, 3619
Zeger, S.L., *see* Diggle, P. 3622
Zeldes, S.P. 3236, 3693
Zellner, A. 3481, 3519, 3541, 3596, 3599
Zellner, A., *see* Abowd, J.M. 3792, 3795
Zhang, J., *see* Arroyo, C.R. 3426
Ziliak, J.P. 3262

SUBJECT INDEX

Accelerated Failure Time (AFT) model 3396
accept–reject method 3575
acceptance rate 3582
adaptive strategy 3663
additively separable functional forms 3693
adverse selection 3283
aggregate fluctuations 3667
aggregate hazard function 3407
aggregate investment functions 3658
aggregate production 3668
aggregate production function 3658
aggregate shocks 3245
aggregate time series 3658
Albert–Chib algorithm 3604
anticipation 3391
aperiodicity 3578
applied econometrics 3665
approximate equilibria 3657
AR(1) process 3242
asset pricing 3670
association measure 3444
asymptotic bias 3197–3200, 3202–3204, 3209
asymptotic refinements 3161–3163, 3171, 3172, 3176, 3180, 3184–3188, 3191–3193, 3195–3197, 3199, 3200, 3202, 3205–3207, 3209–3212, 3216, 3221
asymptotically pivotal 3164, 3177–3180, 3182–3184, 3197, 3207, 3209, 3211, 3213
asymptotically pivotal statistic 3163, 3164, 3175, 3178, 3184–3186, 3188, 3203, 3204, 3209, 3210, 3221
attenuation 3709, 3713, 3763, 3765, 3768, 3785
autocorrelation 3580, 3596, 3659
autoregressive time series 3610
autovariances of real output 3658
auxiliary data 3604

backsolving 3482, 3484, 3486, 3487, 3489
backward pass 3613
balanced growth path 3669, 3698
baseline hazard 3395
basic marginal likelihood identity (BMI) 3629
Bayes factors 3573, 3626
Bayes theorem 3589
Bayesian inference 3267, 3468, 3469, 3473, 3538–3548, 3571
behavioural equilibrium 3664
benchmark data 3656–3659, 3683, 3698
benchmark equilibrium 3655
benchmark equilibrium data sets 3667
benchmark microconsistent data 3660
benchmark year 3665, 3698
benchmarking 3657
best guess approach 3684
best practice 3684, 3685, 3688, 3693, 3699
bias 3161–3163, 3172–3175, 3190, 3198, 3199, 3203, 3219, 3220, 3665
bias-correction methods 3185
bias reduction 3162
binary and ordinal probit 3604
binary choice 3270, 3322, 3323
binary probit 3605
biproportionality constraint 3698
bivariate distribution of unobserved heterogeneity 3442
black boxes 3677
block bootstrap 3189–3191, 3193
blocking scheme 3619
bootstrap 3160–3222
  block 3189–3191, 3193
  sieve 3190, 3191
  stationary 3189, 3190
  wild 3215–3217
bounds 3709, 3721–3725, 3731
Box–Cox model 3279
Brouwer fixed point theorem 3664
burn-in period 3579
business cycles 3655

calibrated models 3655, 3661–3663, 3670, 3671, 3684, 3690, 3694, 3699
calibration 3653–3699
calibrators 3663, 3664, 3685, 3693, 3698
candidate generating density 3581
Carlin–Chib algorithm 3634

cash-in-advance constraints 3671
casual empiricism 3662
censored and truncated regression 3281
censored linear model 3507, 3511, 3512, 3514, 3517
censored regression models 3272
CES (constant elasticity of substitution) 3665
CES functional forms 3665, 3674
choice 3302, 3305, 3306, 3308, 3309
classical measurement error 3710, 3728, 3735, 3770, 3779
co-movements 3659, 3667
Cobb–Douglas functional forms 3665, 3687
coffee table elasticities 3690
cognitive dissonance 3746
cognitive processes 3743, 3744, 3747
cognitive psychology 3743
cointegrated data-generation processes 3195
competing risks 3431
complementarity 3307
components of variance 3232
conditional maximum likelihood 3266
conditional maximum score 3271
conditional systematic sensitivity 3696
conditional systematic sensitivity analysis (CSSA) 3696
conditionally Gaussian state space models 3615
conditioning variables 3693
confidence intervals
  equal-tailed 3183, 3200, 3201
  one-sided 3191, 3193, 3200, 3201, 3203, 3207, 3210
  symmetrical 3178, 3188, 3191–3193, 3200, 3203–3205, 3210
conjugate prior distributions 3637
consistency check 3657
Constant Elasticity of Substitution (CES) 3687
constructive identification 3410
consumption elasticities 3693
control totals 3698
conventional estimation 3698
convergence diagnostics 3596
correlation of the duration variables 3444
correlations 3659
counterfactual analysis 3687
counterfactual change in policies 3656
counterfactual model solution 3695
counterfactual simulation 3695
counterfactual solution 3656
counterfactuals 3685
counting process 3398
covariances of cyclical output 3658
Cramér condition 3177, 3191
cross-equation restrictions 3664, 3665
cross-section econometrics 3232
cross sectional models 3280
crude GMM estimator 3265
curse of dimensionality 3483–3486, 3489, 3495
cyclical component 3670
cyclical variables 3659

d-variate normal distribution 3600
data augmentation 3473, 3501, 3540, 3541, 3589
decompositional analysis 3694
deep parameters 3671, 3686
defective distribution 3421
demand functions 3664
dependence of the two duration variables 3444
deterministic modelling 3655
developed countries 3688
developing countries 3688
deviations from trend 3659
dichotomous choice 3506, 3508, 3511, 3512, 3514, 3515, 3517, 3518, 3522, 3525, 3530, 3533
differential measurement error 3719
dimensionality 3442, 3677
discrete choice 3466, 3467, 3492, 3520–3523, 3528, 3533, 3535–3538
discrete distributions 3416
discretionary time 3669
disequilibrium dynamic process 3664
distortionary taxes 3671
distribution function
  one-sided 3179, 3190
  symmetrical 3178, 3179, 3190
double calibration 3687, 3694, 3695
duration 3745, 3778, 3791, 3799–3801, 3816
duration data 3337
duration dependent 3387
dynamic economy 3698
dynamic expectations models 3332
dynamic general equilibrium macro modellers 3693
dynamic general equilibrium model 3683
dynamic growth model 3678
dynamic models 3331, 3333

dynamic optimization 3467, 3481–3494, 3538, 3543
dynamic stochastic general equilibrium modelling 3662

earnings 3707, 3708, 3710, 3712, 3731, 3733, 3740, 3743, 3747–3770, 3778, 3784, 3790, 3791, 3805, 3827, 3830, 3831, 3833
earnings dynamics 3764, 3765
econometric models 3660, 3663, 3664, 3676
Edgeworth expansion 3171, 3176, 3179, 3199, 3200, 3204, 3206, 3207, 3209, 3210, 3212
education 3707, 3708, 3712, 3718, 3725, 3731, 3733, 3744, 3760, 3763, 3790, 3791, 3804, 3823–3831, 3833
efficient estimation 3259
effluent coefficients 3684
elasticities 3683
elasticities of substitution 3665, 3674
elasticity estimates 3665
elasticity of intertemporal substitution 3693
elasticity parameters 3662, 3676, 3684, 3690
EM algorithm 3639
empirical methodology 3663
endogenous growth models 3671
equal-tailed 3201
  confidence intervals 3183, 3200, 3201
  tests 3180, 3200, 3201
equilibrium 3659, 3664
equilibrium conditions 3660, 3673
equilibrium data 3673, 3686
equilibrium distributions 3598
equilibrium search models 3393
equilibrium solution 3657, 3658, 3660, 3665, 3673, 3675, 3677
equity premium puzzle 3670
error components 3232, 3241
essential survey conditions 3746
estimated models 3655, 3664
estimation 3662–3667
estimation procedures 3684
estimators 3663, 3684
Euler equation 3236, 3238
ex ante equilibria 3694
ex ante policy analysis 3694
ex post analysis 3694
ex post equilibria 3694, 3695
exact equilibria 3657
excess demands 3664
exchangeability 3275
exit rate 3387, 3390
explanatory variables 3663
exponential distributions 3388
exponential family 3266
exponential regression 3251
Extreme Value – Type 1 distribution 3388
extremum estimators 3274

feedback effects 3399
female labour force participation 3257
fertility 3257
finite sample properties 3262
first-differencing 3234, 3430
fixed effects 3232, 3266, 3336, 3430
flexible functional forms 3688
forward pass 3613
Friedmanian/Popperian positivist tradition 3689
fringe benefits 3810, 3811
Frisch elasticity 3692
full conditional distribution 3589
full stationarity 3244
fully articulated models 3663
functional forms 3665, 3666, 3687–3689, 3693, 3699

gamma distribution 3415, 3601
  inverse 3601
Gaussian quadrature 3696
general equilibrium economy 3664
general equilibrium microeconomic models 3655
general equilibrium model 3688, 3698
general equilibrium modelling 3665
general equilibrium parameters 3665
generalized method of moments 3670
genetic determinants 3427
geometric ergodicity 3579
GHK 3502–3505, 3525, 3535, 3561–3563
GHK recursive simulator 3496, 3498, 3499, 3501
Gibbs sampler 3472–3474, 3478, 3479, 3540, 3541
Gibbs sampling method 3589–3595
Gibbs transition kernel 3630
global identification 3321
global interactions 3310
global interactions models 3313
global sensitivity analysis 3696
Global Trade Analysis Project (GTAP) 3683
  data base 3683, 3684
  model 3683, 3684

GLS transformation 3257
GMM (generalized method of moments) estimation 3186–3189, 3191–3193, 3212, 3255–3260, 3262, 3264, 3265, 3431, 3732, 3733, 3829
goodness of fit 3659, 3666
goodness of fit criterion 3663, 3698
Granger's non-causality 3240
growth component 3670

Harberger's model 3657
Hastings–Metropolis algorithm, *see* Metropolis–Hastings algorithm
hazard 3387
hazard function 3387
hazard rate 3387
health 3707, 3717–3721, 3725, 3730, 3733–3735, 3743, 3810–3823, 3831
heterogeneity 3232
heterogeneous AR(1) model 3254
Hicks neutral shock 3658
hidden Markov models 3612
highly persistent data 3247
Hodrick–Prescott filters 3670
homoskedasticity 3243
homothetic preferences 3688
hours worked 3707, 3758, 3759, 3770, 3784–3791, 3831
Huggett model 3678

idealized model 3663
identifiability 3328, 3408
identification 3318–3325, 3327, 3330, 3331, 3338, 3339, 3342, 3343, 3348, 3350, 3361, 3408
identification of individual versus neighborhood effects 3325
identified minimum 3441
identifying restrictions 3667
import and export demand elasticities 3690
importance resampling 3598
importance sampling 3572
incidental parameters 3427
income effects 3688
independence sampling 3470–3472
independent normal innovations 3659
individual heterogeneity 3232
industry 3707, 3708, 3714, 3769, 3802–3805, 3809, 3811
inefficiency factor 3580
information bound 3247, 3260
initial conditions 3244
initial conditions problem 3282
input–output matrices 3698
instrumental variables 3232
instruments 3260
integrated hazard function 3388
integration 3467, 3469, 3470, 3482, 3483, 3485, 3487, 3541
interaction effects 3301
interactions 3300, 3305, 3309, 3347, 3353, 3355–3357, 3413
interactions-based models 3297–3371
intertemporal elasticity 3692
intertemporal substitutability 3659
interval-censored 3405
invariant distribution 3577
inverse gamma distribution 3601
inverse transform method 3573, 3574

job offer 3389
job search models 3389

Kakutani fixed point theorem 3664
Kendall's τ (Kendall's coefficient of concordance) 3446
kernel density 3469
kernel smoothing 3631

labour force participation 3237
labour force status 3791–3800
labour supply elasticities 3693
labour–leisure choice 3658
LAD (least-absolute-deviations) 3176, 3205–3207
lagged dependent variables 3282
lagged-duration dependence 3432
lagged durations 3432
Laplace method 3627
Laplace transform $\mathcal{L}$ 3407
latent linear model 3468, 3506, 3507, 3548, 3555
latent variables 3606
least-absolute-deviations (LAD) 3176, 3205–3207
Leontief functional forms 3665
levels 3259
levels calibration 3686
life-cycle 3678
lifetime distributions 3397
likelihood functions 3664

limited dependent variables models 3232, 3266
limited sensitivity analysis 3696
LIML 3262
Lindholm tradition 3690
Linear Expenditure System (LES) 3687
linear-in-means model 3318, 3322–3324, 3328
linear projection 3234
linear regression 3663
linking densities 3633
local interactions 3312
local interactions models 3313
local reversibility condition 3588
logit model 3267, 3268
Lucas critique 3686

macro calibrations 3662
macro econometric models 3660, 3667
macro model 3693
macro modelling 3660
Mallows metric 3167
marginal/conditional decomposition 3629
marginal likelihood 3468, 3511, 3516–3518, 3525, 3535, 3554, 3561, 3562, 3572, 3626
market activity 3669
market clearing 3664
Markov chain Monte Carlo (MCMC) 3467, 3470, 3472, 3473, 3476, 3479–3481, 3510–3512, 3514–3517, 3521, 3524, 3525, 3529, 3535, 3552–3554, 3559, 3569–3642
Markov chains 3572
  reversible 3578
Markov mixture models 3612
Markov process 3191
Markov switching models 3612
maximum rank correlation 3281
maximum-score (MS) estimator 3205, 3208, 3209, 3270
maximum-score (MS) method 3172
MCEM algorithm 3640
MCMC (Markov chain Monte Carlo), *see* Markov chain Monte Carlo (MCMC)
mean reverting 3713, 3725, 3758
measurement error 3394, 3705–3833
  classical 3708, 3709, 3711–3714, 3716, 3721, 3722, 3725, 3727–3729, 3731, 3737, 3739, 3824, 3828, 3831
  differential 3716–3721, 3734, 3739, 3778
  non-classical 3709, 3739
  nondifferential 3711, 3725, 3727, 3732, 3738
median-regression model 3176, 3184, 3205
method of composition 3576, 3613
method of simulated moments (MSM) 3501, 3502
Metropolis within Gibbs 3479, 3510, 3525, 3560
Metropolis–Hastings algorithm 3473–3476, 3478, 3479, 3571, 3587
  multiple block 3587–3589
Metropolis–Hastings chain 3583, 3585
  random walk 3582
  tailored 3583
Metropolis–Hastings (M–H) method 3580–3589
Metropolis–Hastings probability of move 3631
micro-based modelling 3686
micro-based models 3684
micro calibration 3665, 3673
micro evidence 3693
micro general equilibrium models 3664
micro modellers 3672, 3675, 3689, 3698
micro models 3662
micro policy analyses 3660
microconsistent data sets 3660
microconsistent equilibrium data set 3657
microeconometric parameters 3693
minimum distance 3288
misspecification 3422
Mixed Proportional Hazard (MPH) model 3394
mixture distribution 3617
mixture of exponentials 3421
mixture proposal densities 3598
model calibration 3667
model-dependent approach 3683
model elasticities 3695
model parameterization 3667
model selection 3685
model sensitivity 3696
model solutions 3659, 3660, 3664
model space-parameter space MCMC algorithms 3633
model subsystems 3665
model testing 3685
moments 3660
Monte Carlo methods 3466–3470, 3487, 3511, 3540
moral hazard 3283

MSM (method of simulated moments) 3503–3505
multi-spell data 3426
multinomial logit 3284
multinomial logit model 3269
multinomial probit (MNP) 3468, 3470, 3472, 3494–3505, 3519–3522, 3525, 3537, 3556, 3558
multiple-duration models 3431
multiple individual effects 3252
multiple spells 3426
multiplicative effects 3250
multiplicative models 3251
multivariate binary responses 3622
Multivariate Mixed Proportional Hazard (MMPH) model 3433
multivariate probit 3620
multivariate-*t* distribution 3601

national accounts 3658, 3698
necessities 3688
neoclassical growth 3668
neoclassical growth model 3671
non-replacement subsampling 3170–3172
non-time separable preferences 3658, 3667
nonlinear 3326, 3327, 3340
nonlinear dynamic models 3233
nonlinear-in-means model 3327
nonparametric identification 3408
nonparametric regression model 3280
nonstationarity 3391
normal mixture 3506–3510, 3513, 3514, 3517, 3518, 3520, 3521, 3524, 3525, 3529, 3531, 3535, 3537, 3549–3554, 3556–3558, 3560, 3561
normalization 3263
*n*th-step-ahead transition kernel 3577
numerical general equilibrium models 3658
numerical standard error 3580

observed hazard function 3407
occupation 3707, 3718, 3747, 3802–3805, 3824
omitted variable bias 3663
on-the-job search model 3393
one-factor loading specification 3442
one sector growth model 3658
one-sided 3183, 3201
  confidence intervals 3191, 3193, 3200, 3201, 3203, 3207, 3210
  distribution function 3179, 3190
  hypothesis 3191
  intervals 3203
  tests 3183, 3191, 3193, 3200, 3201, 3207, 3210, 3211
one-tailed tests 3182, 3183, 3213
optimal instruments 3260
ORANI model 3666, 3676
ordinal categorical data 3606
orthogonal deviations 3247, 3256
orthogonality conditions 3235
overlapping generations model 3678
overreporting 3741, 3742, 3745, 3746, 3748, 3758, 3762, 3765–3767, 3771–3777, 3785, 3791, 3811, 3822, 3828, 3831

panel data 3619
panel data models 3229–3290
parameter values 3655
Pareto distribution 3401
partial adjustment 3235
partial equilibrium model 3688
Partial Likelihood 3396
payoff function 3482, 3483, 3490, 3491, 3493, 3494, 3538, 3540, 3543, 3544
perfect international capital mobility 3685
perfect sampling method 3641
permanent component 3658
persistent 3659
π^*-irreducibility 3578
Piggott and Whalley model 3684
pivotal 3165, 3175, 3180, 3213
pivotal statistic 3164, 3213
Poisson regression model 3269
policy evaluation 3685
policy modellers 3672
polynomial approximation 3489–3491, 3493, 3494, 3543, 3548
positive Harris-recurrent 3579
positive (negative) duration dependence 3387
positive stable distributions 3411
posterior density 3571
posterior odds 3626
posterior ordinate 3631
power 3212, 3213
predetermined regressors 3232
predetermined variables 3233
predictability 3398
predictable process 3398
predictive density 3623
prepivoting 3210, 3211
prior density 3571

probability of move 3581
probit 3474, 3492, 3493, 3506, 3507, 3518, 3531, 3548
product characteristics 3661
product of transition kernels 3588
product quality 3661
production functions 3673, 3674
productivity shocks 3658, 3659
profit conditions 3664
Proportional Hazard (PH) model 3395
proposal density 3581
pseudo dominating density 3584
pseudo priors 3633
pure exchange economy 3664

random effects 3267, 3427
random shocks 3660
ranking model 3414
Rao–Blackwellization 3593
RAS (Row and Column Scaling) 3698
real business cycle models 3658, 3665, 3670–3672, 3676, 3689
recall period 3743, 3744, 3760, 3769, 3801, 3804, 3811, 3816, 3821
recursive forward transformation 3260
recursive informational structure 3659
reduced blocking scheme 3614
reduced conditional ordinate 3594, 3629, 3631
reduced-form models 3400
reduced MCMC run 3630
regression parameters 3663
regression with change point 3608
relative price effects 3688
reliability 3719, 3735, 3737, 3763, 3770, 3791, 3817, 3822–3827, 3829, 3832
repeated search 3393
replacement subsampling 3169, 3171
replication check 3660
replication test 3675
reservation wage 3390
reversible jump method 3633
reversible Markov chains 3578
right-censored 3405

sample selection model 3275
sampler performance 3596
Savage–Dickey density ratio 3628
savings elasticity 3685
scale normalizations 3272
scale parameters 3673, 3675
Scarf's algorithm 3657
seam effect 3745
search models 3389
seemingly unrelated regression (SUR) model 3603
selection corrections 3331
selection model 3267, 3467, 3492–3494, 3519, 3521, 3522, 3525, 3530, 3533, 3535–3537
self-evolving organisms 3397
self-selection 3328
self-supplied goods 3684
semiparametric models 3266
sensitivity analysis 3695, 3699
sequential moment restrictions 3255
shocks 3433
sieve bootstrap 3190, 3191
sign of the interaction 3414
simulated and actual moments 3662
simulated annealing method 3639
simulated maximum likelihood (SML) 3496, 3498, 3501, 3505
simulated tempering method 3598
simulation-based inference 3267, 3642
simulation consistent estimate 3631
simulation outcomes 3663
simultaneous estimation 3665
single levels calibration 3690
single-spell data 3405
SLAD (smoothed least-absolute-deviations) 3206, 3207, 3209
2SLS 3262
smooth function model 3163, 3175–3177, 3186, 3199, 3201, 3205, 3210
smoothed least-absolute-deviations (SLAD) 3206, 3207, 3209
smoothed maximum score (SMS) 3208, 3209
smoothness 3670
SMS (smoothed maximum score) 3208, 3209
social desirability 3745, 3746, 3748, 3765, 3766, 3771, 3791, 3802
specification error 3663
state dependence 3283
state space model 3614
stationarity assumption 3390
stationarity restrictions 3248
stationary bootstrap 3189, 3190
stationary distribution 3669
statistical mechanics 3314
statistical tests 3663
steady state 3657, 3660, 3668–3670
steady-state distribution 3657

stochastic disturbances 3658, 3660
stochastic general equilibrium non-monetary macroeconomic models 3655
stochastic growth model 3665
stochastic volatility (SV) model 3616
Stone–Geary functional forms 3687
Stone–Geary/Linear Expenditure System 3687
stratified partial likelihood inference 3426
stratum 3426
strict exogeneity 3233, 3239
strictly exogenous variables 3232
Student-*t* 3507–3512, 3514, 3517, 3520, 3521, 3524, 3528, 3529, 3531, 3533, 3535–3537, 3550, 3552–3554, 3556, 3558, 3560, 3562, 3563
Student-*t* multiple regression 3603
subsampling 3162, 3169–3171, 3190, 3220
successive durations 3437
sufficient statistic 3268
supply correspondences 3664
survey data 3705–3710, 3737, 3741–3743, 3746, 3764, 3779, 3822, 3830–3833
survivor function 3387
symmetrical
  confidence intervals 3178, 3188, 3191–3193, 3200, 3203–3205, 3210
  distribution function 3178, 3179, 3190
  tests 3178, 3180, 3185, 3188, 3191–3193, 3200, 3201, 3203, 3204, 3206, 3207, 3210, 3211, 3217
  two-tailed *t* statistic 3192
  two-tailed tests 3180, 3182, 3183, 3213, 3216
symmetrically normalized 2SLS 3263
systematic part of the hazard. 3395

tail of the unobserved heterogeneity distribution 3409
target density 3582
technological change 3695
technology shocks 3668
telescoping 3744, 3745
tenure 3759, 3763, 3790, 3804–3811
terms of trade 3688
theory with numbers 3656, 3677
time-invariant explanatory variables 3394
time series econometrics 3232
time-varying explanatory variables 3394
Tobit censored regression 3607
Tobit type 3 3277
Tobit type 4 3278
Tobit type 5 3278
Tobit-type models 3272
top – or bottom – coding 3267
trade creation 3672
trade diversion 3672
trade elasticities 3689
trans-log 3688
transfer program 3832
transfer program income 3770–3779
transformation model 3278, 3396
transition kernel 3577
transition probabilities 3232
transition rate from unemployment into employment 3390
transitions 3732, 3778, 3792–3799
transitory component 3659
treatment 3341, 3344
  effects 3340, 3449
truncated 3273
truncated normal distribution 3574
two-person pure exchange economy 3664
two sector model 3685

unconditional systematic sensitivity analysis (USSA) 3696
uncorrelated effects 3238
underground economy 3684, 3687
underidentification 3663
underreporting 3810, 3816, 3821
unemployment 3707, 3710, 3747, 3778, 3784, 3791–3802, 3830
unemployment duration 3389
unfolding brackets 3780
union coverage 3759, 3768, 3805–3811, 3830
union premia 3708
union status 3707, 3809
unit-root 3193, 3194, 3246
univariate regression model 3602
unlimited systematic sensitivity analysis 3696
unobserved explanatory variables 3394
unobserved heterogeneity 3283, 3394
Uruguay Round 3676
utility functions 3673

validation data 3712, 3737–3742, 3748, 3759, 3764, 3766, 3767, 3778, 3780, 3784, 3799, 3824, 3831–3833
value function 3481, 3482, 3484, 3486, 3492, 3538, 3540–3542
values of leisure 3420
variable selection 3636

variable selection in regression 3636
VAT 3684
vector autoregressive process 3659, 3668

wage offer distribution 3389
wages 3707, 3747, 3767, 3769, 3770, 3809, 3810, 3827, 3829
  hourly 3767
Walras Law 3664, 3673
Walrasian general equilibrium 3659
weak exogeneity 3398
weeding out 3407
Weibull duration dependence 3411
weighted constrained quadratic minimization 3698
welfare effects 3671
wild bootstrap 3215–3217
Wishart distribution 3601

HANDBOOKS IN ECONOMICS

1. HANDBOOK OF MATHEMATICAL ECONOMICS (in 4 volumes)
   Volumes 1, 2 and 3 edited by Kenneth J. Arrow and Michael D. Intriligator
   Volume 4 edited by Werner Hildenbrand and Hugo Sonnenschein

2. HANDBOOK OF ECONOMETRICS (in 6 volumes)
   Volumes 1, 2 and 3 edited by Zvi Griliches and Michael D. Intriligator
   Volume 4 edited by Robert F. Engle and Daniel L. McFadden
   Volume 5 edited by James J. Heckman and Edward Leamer
   Volume 6 is in preparation (editors James J. Heckman and Edward Leamer)

3. HANDBOOK OF INTERNATIONAL ECONOMICS (in 3 volumes)
   Volumes 1 and 2 edited by Ronald W. Jones and Peter B. Kenen
   Volume 3 edited by Gene M. Grossman and Kenneth Rogoff

4. HANDBOOK OF PUBLIC ECONOMICS (in 4 volumes)
   Volumes 1 and 2 edited by Alan J. Auerbach and Martin Feldstein
   2 volumes in preparation (editors Alan J. Auerbach and Martin Feldstein)

5. HANDBOOK OF LABOR ECONOMICS (in 5 volumes)
   Volumes 1 and 2 edited by Orley C. Ashenfelter and Richard Layard
   Volumes 3A, 3B and 3C edited by Orley C. Ashenfelter and David Card

6. HANDBOOK OF NATURAL RESOURCE AND ENERGY ECONOMICS (in 3 volumes)
   Edited by Allen V. Kneese and James L. Sweeney

7. HANDBOOK OF REGIONAL AND URBAN ECONOMICS (in 3 volumes)
   Volume 1 edited by Peter Nijkamp
   Volume 2 edited by Edwin S. Mills
   Volume 3 edited by Paul C. Cheshire and Edwin S. Mills

8. HANDBOOK OF MONETARY ECONOMICS (in 2 volumes)
   Edited by Benjamin Friedman and Frank Hahn

9. HANDBOOK OF DEVELOPMENT ECONOMICS (in 4 volumes)
Volumes 1 and 2 edited by Hollis B. Chenery and T.N. Srinivasan
Volumes 3A and 3B edited by Jere Behrman and T.N. Srinivasan

10. HANDBOOK OF INDUSTRIAL ORGANIZATION (in 3 volumes)
Volumes 1 and 2 edited by Richard Schmalensee and Robert R. Willig
Volume 3 is in preparation (editors Mark Armstrong and Robert H. Porter)

11. HANDBOOK OF GAME THEORY with Economic Applications (in 3 volumes)
Volumes 1 and 2 edited by Robert J. Aumann and Sergiu Hart
Volume 3 is in preparation (editors Robert J. Aumann and Sergiu Hart)

12. HANDBOOK OF DEFENSE ECONOMICS (in 1 volume)
Edited by Keith Hartley and Todd Sandler

13. HANDBOOK OF COMPUTATIONAL ECONOMICS (in 1 volume)
Edited by Hans M. Amman, David A. Kendrick and John Rust

14. HANDBOOK OF POPULATION AND FAMILY ECONOMICS (in 2 volumes)
Edited by Mark R. Rosenzweig and Oded Stark

15. HANDBOOK OF MACROECONOMICS (in 3 volumes)
Edited by John B. Taylor and Michael Woodford

16. HANDBOOK OF INCOME DISTRIBUTION (in 1 volume)
Edited by Anthony B. Atkinson and François Bourguignon

17. HANDBOOK OF HEALTH ECONOMICS (in 2 volumes)
Edited by Anthony J. Culyer and Joseph P. Newhouse

18. HANDBOOK OF AGRICULTURAL ECONOMICS (in 3 volumes)
Volumes 1A and 1B edited by Bruce L. Gardner and Gordon C. Rausser
Volume 2 is in preparation (editors Bruce L. Gardner and Gordon C. Rausser)

FORTHCOMING TITLES

HANDBOOK OF SOCIAL CHOICE AND WELFARE ECONOMICS
Editors Kenneth J. Arrow, Amartya K. Sen and Kotaro Suzumura

HANDBOOK OF RESULTS IN EXPERIMENTAL ECONOMICS
Editors Charles Plott and Vernon L. Smith

HANDBOOK OF ENVIRONMENTAL ECONOMICS
Editors Karl-Goran Mäler and Jeff Vincent

HANDBOOK OF THE ECONOMICS OF FINANCE
Editors George M. Constantinides, Milton Harris and René M. Stulz

HANDBOOK ON THE ECONOMICS OF GIVING, RECIPROCITY AND ALTRUISM
Editors Serge-Christophe Kolm and Jean Mercier Ythier

HANDBOOK ON THE ECONOMICS OF ART AND CULTURE
Editors Victor Ginsburgh and David Throsby

HANDBOOK OF ECONOMIC GROWTH
Editors Philippe Aghion and Steven N. Durlauf

All published volumes available