WILEY

Multilevel Statistical Models

4th Edition

Harvey Goldstein

WILEY SERIES IN PROBABILITY AND STATISTICS

Multilevel Statistical Models

WILEY SERIES IN PROBABILITY AND STATISTICS

Established by WALTER A. SHEWHART and SAMUEL S. WILKS

Multilevel Statistical Models

4th Edition

Harvey Goldstein

University of Bristol, UK

WILEY

A John Wiley and Sons, Ltd., Publication

This edition first published 2011

Registered offic
John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom

Library of Congress Cataloguing-in-Publication Data

Goldstein, Harvey.
Multilevel statistical models / Harvey Goldstein. – 4th ed.
p. cm.
Includes bibliographical references and index.
ISBN 978-0-470-74865-7 (cloth)
1. Social sciences–Mathematical models. 2. Social sciences–Research–Methodology. 3. Educational tests and measurements–Mathematical models. I. Title.
H61.25.G65 2010
519.5–dc22

2010023377

A catalogue record for this book is available from the British Library.

Print ISBN: 978-0-470-74865-7
ePDF ISBN: 978-0-470-97340-0
oBook ISBN: 978-0-470-97339-4

Set in 10/12 Times by Aptara.

This book is dedicated to Jon Rasbash who died in March 2010. Without his support, enthusiasm and insight, many of the things discussed in this book would not have happened.

Harvey Goldstein
June 2010

Contents

Preface

In the mid-1980s, a number of researchers began to see how to introduce systematic approaches to the statistical modelling and analysis of hierarchically structured data. The early work of Aitkin *et al*. (1981) on the teaching styles' data and Aitkin's subsequent work with Longford (1987) initiated a series of developments that by the early 1990s had resulted in a core set of established techniques, experience and software packages that could be applied routinely. These methods and further extensions of them are described in this book; they are now applied widely in areas such as education, epidemiology, geography, child growth and household surveys.

In addition to the first, second and third editions of the present text (Goldstein, 1987b, Goldstein, 1995, Goldstein, 2003), several expository volumes have now appeared (see Section 1.15). The present text aims to integrate existing methodological developments within a consistent terminology and notation, provide examples and explain a number of new developments, especially in the areas of latent normal models, missing data, multiple membership structures, errors of measurement and survival data. In almost all cases, these developments are the subject of continuing research.

The main text seeks to avoid undue statistical complexity, with derivations occurring in appendices. Examples and diagrams are used where possible to illustrate the application of the techniques and references are given to other works. The book is intended to be suitable for graduate level courses and as a general reference.

Harvey Goldstein
June 2010

Acknowledgements

This book would not have been possible without the support and dedication of all those who have worked on, or been closely associated with, the Centre for Multilevel Modelling now at the University of Bristol. Those who contributed substantially to the third edition include Jon Rasbash, Min Yang, William Browne, Fiona Steele, Ian Plewis, Michael Healy and Toby Lewis. For the fourth edition, I have continued to benefit considerably from the ideas and advice of Jon Rasbash, William Browne and Fiona Steele. In addition, I have had useful input from James Carpenter, Chris Charlton, Paul Clarke, Bianca DeStavola, Tony Fielding, Kelvyn Jones, Daphne Kounali, George Leckie, Rebecca Pillinger, Tony Robinson and Chris Skinner. All of these people have invested time and effort in correcting mistakes and making many useful suggestions. Many friends and other colleagues, too numerous to mention, have also pointed out errors, suggested improvements and generally provided encouragement. Sincere thanks are due to the Economic and Social Research Council for their almost continuous provision of project funding for methodological developments since 1986. Needless to say, any remaining obscurities or mistakes are entirely my responsibility.

Harvey Goldstein
June 2010

Notation

Examples use 2-level models

Definitio	Symbol
Response variable vector	Y
Explanatory variable design matrix	X
Fixed part explanatory variable design matrix for a single unit	X_{ij} for a level 1 unit X_j for a level 2 unit
Total residuals at each level for a 2-level model	$u_j = \sum_{h=0}^{q_2} u_{hj} z_{hj}^{(2)}$ $e_{ij} = \sum_{h=0}^{q_1} e_{hij} z_{hij}^{(1)}$
Explanatory variable design matrix for level 2 and level 1 random coefficients	$Z^{(2)}$, $Z^{(1)}$
Predicted value from fixed part of model	$\hat{y}_{ij} = X_{ij}\beta = (X\beta)_{ij}$
Raw or total residual for level 1 unit	$\tilde{y}_{ij} = y_{ij} - \hat{y}_{ij}$
Mean raw residual for level 2 unit	$\tilde{y}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} \tilde{y}_{ij}$
Observations (y) independently and identically distributed with specified density function (g)	$y \overset{iid}{\sim} g$
Estimated residual or posterior residual estimate	$\hat{u}_j$, $\hat{e}_{ij}$
Covariance matrix of random coefficients at level i	Ω_i, $\Omega = \{\Omega_i\}$
Parentheses denoting vector or matrix of elements	$\{\ \}$
Covariance matrix of response vector for k-level model	V_k or just V
Contribution to covariance matrix of response vector from level i for k-level model	$V_{k(i)}$, or just V_i
Direct sum of matrices $A_1, \ldots, A_k$	$\bigoplus_{i=1}^{k} A_i$
Kronecker product of matrices A_1, A_2	$A_1 \otimes A_2$
vec operator on matrix A	$vec(A)$
Level 2 cross classification model; classifications indexed by j_1, j_2	$y_{i(j_1 j_2)} = X_{i(j_1 j_2)}\beta + u_{1j_1} + u_{2j_2} + e_{i(j_1 j_2)}$
Multiple membership model; particular case of a level 1 unit belonging to just two level 2 units. j_1, j_2 index units in the same classification.	$y_{i(j_1 j_2)} = X_{i(j_1 j_2)}\beta + w_{1ij_1} u_{j_1} + w_{2ij_2} u_{j_2} + e_{i(j_1 j_2)}$ $w_{1ij_1} + w_{2ij_2} = 1$

A general classificatio notation and diagram

For complex models a more general notation has been used. Response measurements and random effects have a single subscript that identifies the unit. Random effects, and optionally responses, also have a single superscript that identifies the 'classification'. Thus, a variance components two level model, using the example of students and schools, can be written in standard notation as

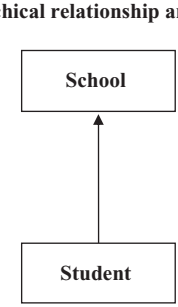
$$y_{ij} = (X\beta)_{ij} + u_j + e_{ij}$$
$$u_j \sim N(0, \sigma_u^2), e_{ij} \sim N(0, \sigma_e^2)$$

and in the general classification notation as

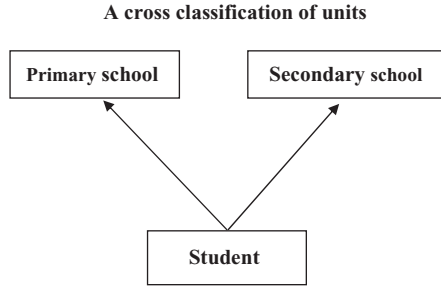
$$y_i^{(1)} = (X\beta)_i + u_{school(i)}^{(2)} + u_{student(i)}^{(1)} \quad school(i) \in (1, \ldots, J)$$
$$student(i) \in (1, \ldots, N) u_{school(i)}^{(2)} \sim N(0, \sigma_{u(2)}^2)$$
$$u_{student(i)}^{(1)} \sim N(0, \sigma_{u(1)}^2) \quad i = 1, \ldots, N$$

By default the lowest level classification is (1) and where no ambiguity arises this superscript may be omitted. The term *school(i)* refers to the school that the *i-th* student belongs to, school being the classification (set of units) with superscript (2). The term *student(i)* refers to student i which is classification (1). Classifications can be general, including nesting relationships, crossings and multiple memberships.

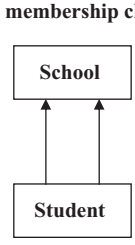
Accompanying this notation is a classification diagram which has the following elements.



Notation diagram 1 Hierarchical relationship among units.



Notation diagram 2 A cross classificatio of units.



Notation diagram 3 A multiple membership classification

Glossary

Cluster	A grouping containing 'lower level' elements. For example in a sample survey the set of households in a neighbourhood.
Cross classification	A structure where lower level units are grouped within the cells of a multiway classification of higher level units
Design matrix	In the fixed part of the model, the matrix of values of the explanatory variables X. In the random part the matrix of explanatory variables Z.
Explanatory variable	Also known as an 'independent' variable. In the fixed part of the model usually denoted by x and in the random part by z.
Fixed part	That part of a model represented by $X\beta$, that is the average relationship. The parameters, β, are referred to as 'fixed parameters'.
Level	A component of a data hierarchy. Level 1 is the lowest level, for example students within schools or repeated measurement occasions within individual subjects.
Level $\boldsymbol{n}$ variation	The variation among level $\boldsymbol{n}$ unit measurements.
Multiple membership	A structure where a level unit may be nested within one or more higher level units.
Nesting	The clustering of units into a hierarchy
Random part	That part of a model represented by Zu, that is the contribution of the random variables u, at each level. The parameters associated with the random variables, i.e. variances and covariances are referred to as 'random parameters'.
Response variable	Also known as a 'dependent' variable. Denoted by y.
Unit	An entity defined at a level of a data hierarchy. For example an individual student will be a level 1 unit within a level 2 unit such as a school.

1

An introduction to multilevel models

1.1 Hierarchically structured data

Many kinds of data, including observational data collected in the human and biological sciences, have a *hierarchical, nested,* or *clustered* structure. For example, animal and human studies of inheritance deal with a natural hierarchy where offspring are grouped within families. Offspring from the same parents tend to be more alike in their physical and mental characteristics than individuals chosen at random from the population at large. For instance, children from the same family may all tend to be small, perhaps because their parents are small or because of a common impoverished environment. Many designed experiments, such as clinical trials carried out in several randomly chosen centres or groups of individuals, also create data hierarchies.

For now, we are concerned only with the *fact* of such hierarchies, not their provenance. The principal applications are those from the social and medical sciences, but the techniques are, of course, applicable more generally. In subsequent chapters, as we develop the theory and techniques with examples, we see how a proper recognition of these natural hierarchies allows us to obtain more satisfactory answers to important questions.

We refer to a hierarchy as consisting of *units* grouped at different *levels*. Thus offspring may be the level 1 units in a 2-level structure where the level 2 units are the families: students may be the level 1 units clustered or nested within schools that are the level 2 units.

The existence of such data hierarchies is neither accidental nor ignorable. Individual people differ, as do individual animals, and this differentiation is mirrored in all kinds of social activity where the latter is often a direct result of the former; for

Multilevel Statistical Models: 4th Edition Harvey Goldstein

example, when students with similar motivations or aptitudes are grouped in highly selective schools or colleges. In other cases, the groupings may arise for reasons less strongly associated with the characteristics of individuals, such as the allocation of young children to elementary schools, or the allocation of patients to different clinics. Once groupings are established, even if their establishment is effectively random, often they will tend to become differentiated. This differentiation implies that the group and its members both influence and are influenced by the group membership. To ignore this risks overlooking the importance of group effects, and may also render invalid many of the traditional statistical analysis techniques used for studying data relationships.

We look at this issue of statistical validity in the next chapter. For now, one simple example will show its importance. A well-known and influential study of the teaching styles used with primary (elementary) school children carried out in the 1970s (Bennett, 1976), claimed that children exposed to so-called 'formal' styles of teaching reading exhibited more progress than those who were not. The data were analysed using traditional multiple regression techniques which recognised only the individual children as the units of analysis and ignored their groupings within teachers and into classes. The results showed statistically significant differences. Subsequently, Aitkin *et al*. (1981) demonstrated that when the analysis accounted properly for the grouping of children into classes, the significant differences disappeared and the 'formally' taught children could not be shown to differ from the others.

This re-analysis is the first important example of a *multilevel* analysis of social science data. In essence what was occurring here was that the children within any one classroom, because they were taught together, tended to be similar in their performance. As a result they provided rather less information than would have been the case if the same number of students had been taught separately by different teachers. In other words, the basic unit for purposes of comparison should have been the teacher not the student. The function of the students can be seen as providing, for each teacher, an estimate of that teacher's effectiveness. Increasing the number of students per teacher would increase the precision of those estimates but not change the number of teachers being compared. Beyond a certain point, simply increasing the numbers of students in this way hardly improves things at all. On the other hand, increasing the number of teachers to be compared with the same or an even smaller number of students per teacher considerably improves the precision of the comparisons.

Researchers have long recognised this issue. In education, there has been much debate (see Burstein *et al*., 1980) about the so-called 'unit of analysis' problem just outlined. Before multilevel modelling became well developed as a research tool, the problems of ignoring hierarchical structures were reasonably well understood, but they were difficult to solve because powerful general purpose tools were unavailable. Special purpose software, for example, for the analysis of genetic data, has been available longer but this was restricted to 'variance components' models (see Chapter 2) and was not suitable for handling general linear models. Sample survey workers have recognised this issue in another form. When population surveys are carried out, the sample design typically mirrors the hierarchical population structure, in terms of geography and household membership. Elaborate procedures have been developed

to take such structures into account when carrying out statistical analyses. We look at this in more detail in Chapter 10.

The remainder of this chapter discusses some general issues and introduces the major topics explored in this book.

1.2 School effectiveness

Schooling systems present an obvious example of a hierarchical structure, with pupils clustered within schools, which themselves may be clustered within education authorities or boards. It therefore provides a useful way to introduce some basic ideas of multilevel modelling. Educational researchers have long been interested in comparing schools and other educational institutions, most often in terms of the achievements of their pupils. Such comparisons have several aims, including the aim of public accountability (Goldstein, 1997) but, in research terms, interest is usually focused upon studying the factors that explain school differences.

Consider the common example where test or examination results at the end of a period of schooling are collected from a randomly chosen sample of schools. The researcher wants to know whether a particular kind of subject streaming practice in some schools is associated with improved examination performance. She also has good measures of the pupils' achievements when they started the period of schooling so that she can control for this in the analysis. The traditional approach to the analysis of these data would be to carry out a regression analysis, using performance score as the response, to study the relationship with streaming practice, adjusting for the initial achievements as covariates. This is very similar to the teaching styles analysis described in the previous section, and suffers from the same lack of validity through failing to take account of the school level clustering of students.

An analysis that explicitly models the manner in which students are grouped within schools has several advantages. Firstly, it enables data analysts to obtain statistically efficient estimates of regression coefficients. Secondly, by using the clustering information it provides correct standard errors, confidence intervals and significance tests, and these generally will be more 'conservative' than the traditional ones that are obtained simply by ignoring the presence of clustering – just as Bennett's previously statistically significant results became non-significant on reanalysis. Thirdly, by allowing the use of covariates measured at any of the levels of a hierarchy, it enables the researcher to explore the extent to which differences in average examination results between schools are accountable for by factors such as organisational practice or in terms of other characteristics of the students. It also makes it possible to study the extent to which schools differ for different kinds of students, for example to see whether the variation between schools is greater for initially high scoring students than for initially low scoring students (Goldstein *et al.*, 1993) and whether some factors are better at accounting for or 'explaining' the variation for the former students than for the latter. Finally, there may be interest in the relative ranking of schools, using the performances of their students after adjusting for intake achievements. This can be

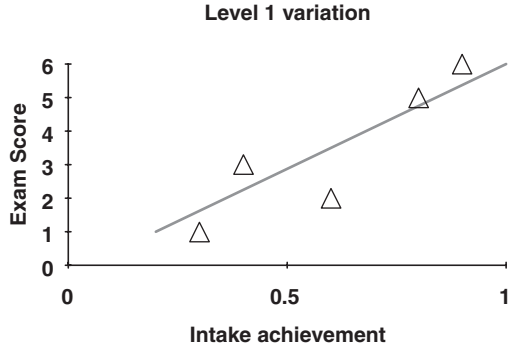


Figure 1.1 A simple regression.

done straightforwardly using a multilevel modelling approach and we shall see an example in Chapter 2.

To fix the basic notion of a level and a unit, consider Figures 1.1 and 1.2, which are based on hypothetical relationships.

Figure 1.1 shows the exam score and intake achievement scores for five students in a school, together with a simple regression line fitted to the data points. The residual variation in the exam scores about this line is the *level 1 residual variation*, since it relates to level 1 units (students) within a sample level 2 unit (school). In Figure 1.2 the three lines are the simple regression lines for three schools, with the individual student data points removed. These vary in both their slopes and their intercepts (where they would cross the exam axis), and this variation is the *level 2 variation*. It is an example of complex level 2 variation since both the intercept and slope parameters vary.

The other extreme to an analysis which ignores the hierarchical structure is one which treats each school completely separately by fitting a different regression model within each one. In some circumstances, for example where we have very few schools

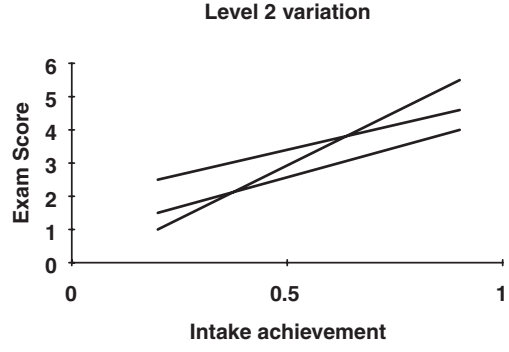


Figure 1.2 Complex level 2 variation.

and moderately large numbers of students in each, this may be efficient. It may also be appropriate if we are interested in making inferences about just those schools. If, however, we regard these schools as a (random) sample from a population of schools and we wish to make inferences about the variation between schools in general, then a full multilevel approach is called for. Likewise, if some of our schools have very few students, fitting a separate model for each of these will not yield reliable estimates: we can obtain more precision by regarding the schools as a sample from a population and using the information available from the whole sample data when making estimates for any one school. This approach is especially important in the case of repeated measures data where we typically have very few level 1 units per level 2 unit (Chapter 5).

We introduce the basic procedures for fitting multilevel models to hierarchically structured data in Chapter 2 and discuss the design problem of choosing the numbers of units at each level in Chapter 3.

1.3 Sample survey methods

We have already mentioned sample survey data. The standard literature on surveys, reflected in survey practice, recognises the importance of taking account of the clustering in complex sample designs. Thus, in a household survey, the first stage sampling unit will often be a well-defined geographical unit. From those which are randomly chosen, further stages of random selection are carried out until the final households are selected. Because of the geographical clustering exhibited by measures such as political attitudes, special procedures have been developed to produce valid statistical inferences, for example, when comparing mean values or fitting regression models (Skinner *et al*., 1989).

While such procedures usually have been regarded as necessary they have not generally merited serious substantive interest. In other words, the population structure, insofar as it is mirrored in the sampling design, is seen as a 'nuisance factor'. By contrast, the multilevel modelling approach views the population structure as of potential interest in itself, so that a sample designed to reflect that structure is not merely a matter of saving costs as in traditional sample design, but can be used to collect and analyse data about the higher level units in the population.

Although the direct modelling of clustered data is statistically efficient, it will generally be important to incorporate weightings in the analysis which reflect the sample design or, for example, patterns of non-response, so that robust population estimates can be obtained and so that there will be some protection against serious model misspecification. A procedure for introducing external unit weights into a multilevel analysis is discussed in Chapter 3 and survey data are discussed in Chapter 10.

1.4 Repeated measures data

A different example of hierarchically structured data occurs when the same individuals or units are measured on more than one occasion, as occurs in studies of animal

and human growth. Here the occasions are clustered within individuals that represent the level 2 units with measurement occasions as the level 1 units. Such structures are typically strong hierarchies because there is much more variation between individuals in general than between occasions within individuals. In the case of child height growth, once we have adjusted for the overall trend with age, the variance between successive measurements on the same individual is generally no more than about 5 % of the variation in height between children.

The traditional literature on the analysis of such repeated measurement data (see, for example, Goldstein, 1979), has more or less successfully confronted the statistical problems. It has done so, however, by requiring that the data conform to a particular, balanced, structure. Broadly speaking these procedures require that the measurement occasions are the same for each individual. This may be possible to arrange, but often in practice individuals will be measured irregularly, some of them a great number of times and some perhaps only once. By considering such data as a 2-level structure, however, we can apply the standard set of multilevel modelling techniques that allow for any pattern of measurements while providing statistically efficient parameter estimation. At the same time, modelling such data as a 2-level structure presents a simpler conceptual understanding and leads to a number of interesting extensions (see Chapter 5).

One particularly important extension occurs in the study of growth where the aim is to fit growth curves to measurements over time. In a multilevel framework this involves, in the simplest case, each individual having their own straight line growth trajectory with the intercept and slope coefficients varying between individuals (level 2). When the level 1 measurements, considered as deviations from each individual's fitted growth curve, are not independent but have an autocorrelated or time series structure, neither the traditional procedures nor the basic multilevel ones are adequate. This situation may occur when measurements are made very close together in time so that a 'positive' deviation from the curve at one time implies also a positive deviation after the short interval before the next measurement. Chapter 5 considers methods for handling such data.

1.5 Event history and survival models

Modelling the time spent in various states or situations is important in a number of areas. In industry the 'time to failure' of components is a key factor in quality control. In medicine, the survival time is a fundamental measurement in studying certain diseases. In economics, the duration of employment periods is of great interest. In education, researchers often study the time students spend on different tasks or activities.

In studying employment histories, any one individual will generally pass through several periods of employment or unemployment, while at the same time changing his characteristics, for example his level of qualifications. From a modelling point of view we need to consider the length of time spent in each type of employment, relating this both to constant factors such as an individual's social origins or gender,

and to changing or time dependent factors such as qualifications and age. In this case the multilevel structure is analogous to that for repeated measures data, with periods taking the place of occasions. Furthermore, generally we would have a further, higher level of the hierarchy, since individuals, which are the level 2 units, are themselves typically clustered into workplaces, which now constitute level 3 units.[1] The structure may be even more complicated if these workplaces change from period to period; to include this level in our model, we need to consider cross-classifications of the units (see below). There are particular problems that arise when studying event duration data that are encountered when some information is 'censored' in the sense that instead of being able to observe the actual duration we only know that it is longer than some particular value, or in some cases less than a particular value (see Chapter 11).

1.6 Discrete response data

Until now, we have assumed implicitly that our response or dependent variable is continuously distributed; for example, an exam score or anthropometric measure such as height. Many kinds of statistical models, however, deal with categorised responses, in the simplest case with proportions. Thus, we might be interested in a mortality rate, or an examination pass rate and how these vary from area to area or from school to school.

In studying mortality rates in a population, it is often of great concern to try to understand the factors associated with variations from area to area or community to community. This produces a basic 2-level structure with individuals at level 1 and communities at level 2. A typical study might record deaths over a given time period together with the characteristics of the individuals concerned, and level 2 characteristics of the communities, such as their sizes or social compositions. One analysis of interest would be to see whether any of these explanatory variables could explain between-community variation. Another interest might be in studying whether mortality rate differences, say between men and women, varied from community to community.

Such models, part of the class known as generalised linear models, have been available for some time for single level data (McCullagh and Nelder, 1989), with associated software. In Chapter 4 we show how to fit multilevel models with different types of categorical response. Chapter 7 extends this to consider multivariate models with mixtures of different response types.

1.7 Multivariate models

An interesting special case of a 2-level model is the multivariate linear (or generalised linear) model. Suppose we have taken several measurements on an individual, for

[1] Formally, we can regard unemployment for this purpose as a particular workplace.

example their systolic and diastolic blood pressure and their heart rate. If we wish to analyse these together as response variables we can do so by setting up a multivariate, in this case 3-variate, model with explanatory variables such as age, gender, social background, smoking exposure, etc. We can think of this as a 2-level model by considering each individual as a level 2 unit, with the three measurements constituting the level 1 units, rather as occasions did for the repeated measures model. Chapter 6 shows how this formal device for specifying a multivariate model yields considerable benefits. By considering further higher levels, such as clinics, we have a simple way of specifying a multivariate multilevel model and we can also have models where the responses can be measured at different levels of a data hierarchy; see Chapter 5 for an example. Also, if some individuals do not have all the measurements, for example if they are randomly missing a blood pressure measurement, then this is automatically taken account of in the analysis, without the need for special procedures for handling missing data.

A particularly important application occurs where measurements are missing by design rather than at random. In certain kinds of surveys, known as rotation designs, and in certain kinds of educational assessments known as matrix sample designs, each individual unit has only a subset of measurements made on it. For example, in large-scale testing programmes, the full range of tests may be too extensive for any one student, so that each student responds to only one, randomly assigned, combination. Such designs can be viewed as having a multivariate response, with the full set of tests constituting the complete multivariate response vector, and every student having some tests missing. Such designs can become rather complex, especially since the students themselves are clustered into schools. By viewing the data as a single hierarchy in which the multivariate responses are level 1, we obtain an efficient and readily interpretable analysis.

The multivariate multilevel model is also used as the basis for dealing with missing data in multilevel models and this is developed in Chapter 16.

1.8 Nonlinear models

Some kinds of data are better represented in terms of nonlinear rather than linear models. For example, the modelling of discrete response data is considered formally as a case of modelling nonlinear data. Many kinds of growth data are conveniently modelled in this way, especially during periods of rapid and complex growth such as early infancy and at the approach to adulthood when growth approaches an upper asymptote (Goldstein, 1979). Other examples arise when the response variable has inherent constraints. For example, biochemical activity patterns in patients may exhibit asymptotic behaviour, or cyclical patterns, both of which may be difficult to model using purely linear models. Chapter 9 introduces such models and shows how to extend the linear multilevel model to this case. It also considers cases where variances and covariances can be modelled as nonlinear functions of explanatory variables (see also Chapter 17).

1.9 Measurement errors

Many variables of interest in the human sciences contain 'noise' or random measurement error. This may be due to observer error as when measuring the weight of an animal, or an inherent result of being able to measure only a small sample of behaviour as in educational testing. It is well known that when variables in statistical models contain relatively large components of such error the resulting statistical inferences can be very misleading unless careful adjustments are made (Fuller, 2006). In the case of simple regression, when the explanatory or independent variable is measured with error, the usual estimate of the regression line slope is an underestimate compared to that which would result if the measurement were available without error. This is particularly important in studies of school effectiveness where the fitting of intake achievement scores is important but where such scores often have large components of measurement error.

An important case is where we have a level 2 variable that is a 'compositional' variable. That is, it is a measurement aggregated from the characteristics of the level 1 units within the level 2 units. Thus, for example, the mean intake achievement and the standard deviation of the intake achievements of all the pupils in a school are compositional variables that may, and indeed sometimes do, affect the final achievements of each individual student. Likewise, in a household survey, we may consider that a measure of the average social status or the percentages of households in each social group, using all the households in the immediate community, are important explanatory variables to fit in a model. The problem arises when it is possible to collect data only upon some of the level 1 units, this often being the case with household sample surveys. What we then have is an estimate of a compositional variable that is measured with error, in the case of household surveys typically with a very large error. In many educational studies, this also occurs where only a small proportion of students within a class or school are sampled. Chapter 14 discusses the problems of dealing with measurement errors in multilevel models.

1.10 Cross classification and multiple membership structures

We have already alluded to examples where units are cross-classified as well as clustered. In area studies the definition of an individual's geographical area is contingent upon the context being considered. Thus, the relevant location unit for purposes of leisure may not be the same as that surrounding the environment of work or schooling. We can conceive, formally, of individuals belonging simultaneously to both types of unit, each of which may have an influence on a person's life.

In most schooling systems, students move from elementary to secondary or high school. We might expect that both the elementary and secondary schools attended will influence a student's achievements or attitudes measured at the end of secondary school. Thus the level 2 units are of two types, elementary school and secondary

school, where each 'cell' of their cross classification contains some, or possibly no students. In this example, a third classification could be the area or neighbourhood where the student lives. Chapter 12 explores such cross-classified structures.

An interesting situation occurs where for a single level 2 classification, level 1 units may belong to more than one level 2 unit. A sociological example concerns childrens' and adults' friendship patterns. An individual may belong to several 'friendship groups' simultaneously. The characteristics of the members of each group will influence such an individual, in relation to the individual's exposure to the group. In a longitudinal study of schooling, many students will change schools during the course of the study. The contribution to the response from schools will therefore reflect, for these students, the 'effect' of every school they have attended. With a suitable set of weights to reflect the time spent in each school this can be taken into account in the analysis. Such 'multiple membership models' are discussed in Chapter 13.

To handle the complexity of multiple membership and cross-classified structures, as well as mixtures of these, a special notation and set of diagrams will be introduced that allows a complete specification of such models (see Notation).

1.11 Factor analysis and structural equation models

In many areas of the social sciences where measurements are difficult to define precisely, an investigator might suppose that there is some underlying construct which cannot be measured directly but nevertheless can be assessed indirectly by measuring a number of relevant indicators. Structural equation modelling, and in particular the special case of factor analysis, was developed for this purpose, typically dealing with individuals' behaviour, attitudes or mental performance. Where individuals are grouped within hierarchies, for the reasons already discussed, it is important to carry out such analyses in a multilevel framework. For example, we may be interested in underlying individual attitudes based upon a number of indicators. Data on such indicators may be available over time and we can postulate a model whereby the underlying attitude varies from individual to individual (level 2) and also varies randomly over time within individuals (level 1). The model can then be further elaborated by studying whether there is any systematic change over time and whether this varies across individuals. Chapter 8 discusses such models.

1.12 Levels of aggregation and ecological fallacies

When studying relationships among variables, there has often been controversy about the appropriate 'unit of analysis'. We have alluded to this already in the context of ignoring hierarchical data clustering and, as we have seen, the issue is resolved by explicit hierarchical modelling.

One of the best known early illustrations of what is often known as the ecological or aggregation fallacy was the study by Robinson (1950) of the relationship between literacy and ethnic background in the United States. When the mean literacy rates

and mean proportions of Black Americans for each of nine census divisions are correlated the resulting value is 0.95, whereas the individual-level correlation ignoring the grouping is 0.20. Robinson was concerned to point out that aggregate-level relationships could not be used as estimates for the corresponding individual-level relationships and this point is now well understood. In Chapter 3, we discuss some of the statistical consequences of modelling only at the aggregate level.

Sometimes the aggregate level is the principal level of interest, but nevertheless a multilevel perspective is useful. Consider the example (Derbyshire, 1987) of predicting the proportion of children socially 'at risk' in each local administrative area for the purpose of allocating central government expenditure on social services. Survey data are available for individual children with information on risk status so that a prediction can be made using area based variables as well as child and household based variables. The probability (π) of a child being 'at risk' was estimated by the following (single level) equation

$$\text{logit}(\pi) = -6.3 + 5.9x_1 + 2.2\,x_2 + 1.5x_3$$

where x_1 is the proportion of children in the area in households with a lone parent, x_2 is the proportion of households in each area which have a density of more than 1.5 persons per room and x_3 is the proportion of households whose 'head' was born in the British 'New Commonwealth' or Pakistan. All these explanatory variables are measured at the aggregate area level and the response is the proportion of children at risk in each area. Although we can regard this analysis as taking place entirely at the area level (with suitable weighting for the number of children in each area), there are advantages in thinking of it as a 2-level model with each child being a level 1 unit and the response variable being the binary response of whether or not the child is at risk.

Firstly, this allows us to incorporate possibly important variables that are measured at the child level, for example whether or not each child's household is overcrowded. Including such level 1 variables may greatly improve the predictive power of the model. With the results of such a model we can then form a prediction for each area by aggregating over the known numbers of children living in overcrowded households. Secondly, the possibility of modelling the characteristics of children or their households allows the possibility of an allocation formula that can take account of costs and benefits related to the actual composition of each area in terms of these child characteristics.

1.13 Causality

In the natural sciences, experimentation has a dominant position when making causal inferences. This is both because the units of interest can be manipulated experimentally, typically using random allocation, and because there is a widespread acceptance that the results of experiments are generalisable over space and time. The models described in this book can be applied to experimental or non-experimental data, but

the final causal inferences can differ. Nevertheless, most of the examples used are from non-experimental studies in the human sciences and a few words on causal inferences from such data may be useful.

If we wish to answer questions about a possible causal relationship between, say, class size and educational achievement, an experimental study would need to assign different numbers of level 1 units (students) randomly to level 2 units (classes or teachers) and study the results over a time period of several years. This would be time consuming and could create ethical problems. In addition to such practical problems, any single study would be limited in time and place, and require extensive replication before results could be generalised confidently. The specific context of any study is important; for example, the state of the educational system and the resources available at the time of the study. The difficulty from an experimental viewpoint is that it is practically impossible to allocate randomly with respect to all such possible confounding factors.

A further limitation of randomised controlled trails (RCTs) is that they cannot necessarily deal with situations where the *composition* of a higher level unit interacts with the treatment of interest, to affect the responses of lower level units. Thus, in schooling studies the size of class may affect the progress of students only when the proportion of 'low achieving' students is above a certain threshold. Randomisation will tend to eliminate classes with extreme proportions so that such effects may not be discovered. Goldstein (1998) looks at this case in more detail.

None of this is to say that randomised experiments should never be undertaken, rather that on their own they may have limited potential for making general statements about causality. Whether an experiment fails or succeeds in demonstrating a relationship, there will almost always be further explanations for the findings which require study. Even if an experiment appears to eliminate a possible relationship, for example, demonstrating a negligible relationship between class size and attainment, it may be legitimate to query whether a relationship nevertheless exists for specific subgroups of the population.

In the pursuit of causal explanations, it is desirable to have some guiding underlying principles or theories. It is these which will tell us what kinds of things to measure and how to be critical of findings. For example, in studies of the relationship between perinatal mortality and maternal smoking in pregnancy (Goldstein, 1976) we can attempt to adjust for confounding factors, such as poverty, which may be responsible for influencing both smoking habits and mortality. We can also study how the relationship varies across groups and seek measures which explain such variation. We might also, in some circumstances, be able to carry out randomised experiments, assigning, for example, intensive health education to a randomly selected 'treatment' group and comparing mortality rates with a 'control' group.

A multilevel approach could be useful here in two different ways. Firstly, pregnant women will be grouped hierarchically, geographically and by medical institution, and the between-area and between-institution variation may affect mortality and the relationship between mortality and smoking. Secondly, we will be able often to obtain serial measurements of smoking, so allowing the kind of repeated measures

2-level modelling discussed earlier. This will allow us to study how changes in smoking are related to mortality, and permit a more detailed exploration of possible causal mechanisms.

Multilevel models can often be used to identify units with extreme values. In school effectiveness studies, an exploration of school-level residual estimates (see Chapter 2) may identify those which are highly atypical, having adjusted for 'contextual' variables such as the intake characteristics of their students. These can then be selected for further scrutiny, for example by means of intensive case studies, so forming a link between the quantitatively based multilevel analysis and a more qualitatively based investigation which would seek to identify detailed causal processes.

The notion of causation, especially in non-randomised studies, is controversial and has a long philosophical history. Recent work has extended the range of tools available for studying causality and a useful introduction is given by Sobel (2000); Rosenbaum (1995) provides a detailed discussion of issues. A particular assumption in deriving causal inferences that is important in multilevel modelling, is the 'stable unit treatment assumption' (SUTVA). This states that the response to a 'treatment' assigned to an individual, for example being placed in a small rather than large class for teaching purposes, does not depend on the assignments to other individuals – that there is no interference between units. Where there are hierarchical structures, such as schools, within which different treatments may occur, this 'non-interference' assumption may not hold. It may be possible to model such dependencies or to redesign studies to avoid this problem. A discussion of this problem in the context of class size studies is given by Blatchford *et al*. (1998).

Finally, many of the concerns addressed by multilevel models are to do with straightforward prediction. Thus, for example, in Chapter 6 we use a 2-level model of children's growth for the purpose of predicting adult height. In studies of school effectiveness we may be interested in understanding the causes of school differences, but we may be concerned also with the less ambitious task of predicting which school is likely to produce the best (on average) examination result for a student with given initial characteristics and achievements.

1.14 The latent normal transformation and missing data

In Chapter 7, we show how various discrete responses can be modelled simultaneously by transforming them jointly to an underlying multivariate normal distribution. This has an important application to ways of handling missing data, as discussed in Chapter 16. In particular, Chapter 16 shows how a multiple imputation approach can incorporate not only mixtures of different types of response but also responses at different levels of a data hierarchy. This provides a very general procedure for handling missingness in complex data structures.

1.15 Other texts

While the present volume aims to provide a comprehensive coverage of the topic of multilevel models, there are now many other texts which deal with specialised areas. Many of these are referenced in the appropriate chapters, but there are also several books which provide good introductions as well as detailed worked examples and technical details. Among these are Snijders and Bosker (1999), Little *et al*. (2000), Heck and Thomas (2000), McCulloch and Searle (2001), Hox (2002), Bryk and Raudenbush (2002), Skrondal and Rabe-Hesketh (2004), Lee *et al*. (2006) and Gelman and Hill (2007). There are also edited collections of articles on particular application areas. Leyland and Goldstein (2001) bring together a collection of papers on the multilevel modelling of health statistics and Courgeau (2007) brings together a number of perspectives on the interpretation of multilevel data.

1.16 A caveat

The purpose of this book is to bring together techniques for the analysis of highly structured multilevel data. The application of such techniques has already begun to yield new and important insights in a number of areas as the following chapters illustrate. Software is now widely available (Chapter 18), so that the application of these techniques should become relatively straightforward, even routine.

All this is welcome, yet despite their usefulness, models for multilevel analysis cannot be a universal panacea. In circumstances where there is little structural complexity, they may be hardly necessary and traditional single level models may suffice both for analysis and presentation. On the other hand multilevel analyses can bring extra precision to attempts to understand causality, for example, by making efficient use of student achievement data in attempts to understand differences between schools. They are not, however, substitutes for well grounded substantive theories, nor do they replace the need for careful thought about the purpose of any statistical modelling. Furthermore, by introducing more complexity they can extend but not necessarily simplify interpretations.

Multilevel models are tools to be used with care and understanding.

2

The 2-level model

2.1 Introduction

We now introduce the 2-level model together with the basic notation which we shall use and develop throughout the book. We look at alternative ways of setting up and motivating the model, introducing procedures for estimating parameters, forming and testing functions of the parameters and constructing confidence intervals. We introduce alternative methods of estimation which will be used and elaborated upon in subsequent chapters.

To make matters concrete consider the following dataset, one we shall use several times: it consists of 728 pupils in 48 primary (elementary) schools in inner London, part of the 'Junior School Project' (JSP). We consider two measurement occasions: the first when the pupils were in their fourth year of schooling, that is the year they attained their eighth birthday, and three years later in their final year of primary school. Our data are in fact a subsample from a more extensive dataset, described in detail in Mortimore *et al*. (1988). We use the scores from mathematics tests administered on these two occasions together with information collected on the social background of the pupils and their gender. In this chapter the data are used primarily to illustrate the development of basic 2-level modelling. In Chapter 3, we shall be studying more elaborate models which will enable us to handle these data more efficiently.

Figure 2.1 is a scatterplot of the mathematics test score at age 11 by the test score at age 8. In this plot no distinction is made between the schools to which the pupils belong. Notice that there is a general trend, with increasing 8-year scores associated with increasing 11-year scores. Notice also the narrowing of the between pupil variation in the age 11 score with increasing age 8 score; an issue to which we shall return.

In Figure 2.2, the scores within two particularly different schools have been selected, represented by different symbols. Two things are apparent immediately. The

Multilevel Statistical Models: 4th Edition Harvey Goldstein

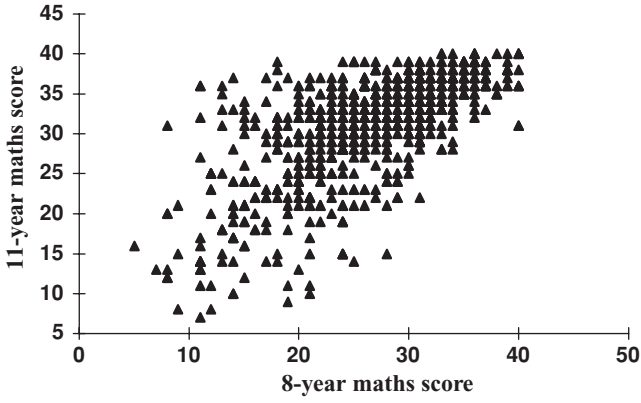


Figure 2.1 Scatterplot of 11-year by 8-year mathematics test scores. Some points represent more than one pupil.

school represented by the circles shows a steeper 'slope' than the school represented by the filled triangles and for most age 8 scores, the age 11 scores for this school tend to be lower than for the other school. Both these features are now addressed by formally modelling these relationships.

Consider first a simple model for one school, relating age-11 score to age-8 score. We write

$$y_i = \alpha + \beta\, x_i + e_i \tag{2.1}$$

where i indexes the individual student and standard interpretations can be given to the intercept (α), slope (β) and residual (e_i). We would also typically assume that the residuals follow a normal distribution with a zero mean and common variance, that

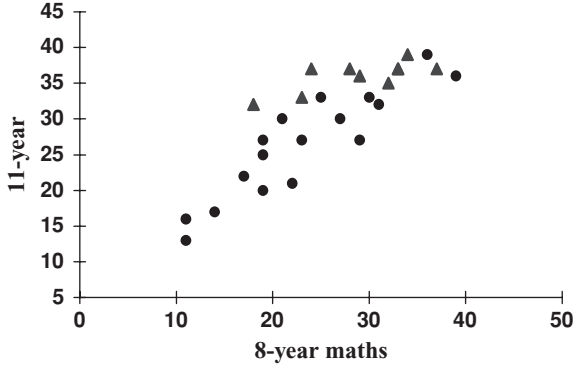


Figure 2.2 Scatterplot of 11-year by 8-year mathematics test scores for two schools.

is, $e_i \sim N(0, \sigma_e^2)$, but for now we shall concentrate on the variance properties of the residuals. We follow the normal convention of using Greek letters for the regression coefficients and place a circumflex over any coefficient (parameter) which is a sample estimate. This is the formal model for Figure 1.1 in the previous chapter and describes a single-level relationship. To describe simultaneously the relationships for several schools we write, for each school *j,*

$$y_{ij} = \alpha_j + \beta_j x_{ij} + e_{ij} \quad e_{ij} \sim N(0, \sigma_e^2) \tag{2.2}$$

This is now a formal model describing Figure 1.2 where *j* refers to the level 2 unit (school) and again *i* to the level 1 unit (pupil).

As it stands, (2.2) is still essentially a single level model, albeit describing a separate relationship for each of the *m* schools. In some situations, for example, where there are few schools and interest centres on just those schools in the sample, we may analyse (2.2) by fitting all the $2m + 1$ parameters, namely

$$(\alpha_j, \beta_j) \quad j = 1, \ldots, m \quad \sigma_e^2$$

assuming a common 'within-school' residual variance and separate lines for each school.

If we wish to focus not just on these schools, but on a wider 'population' of schools, then we have to regard the chosen schools as giving us information about the characteristics of all the schools in the population. Just as we choose random samples of individuals to provide estimates of population means etc., so a randomly chosen sample of schools can provide information about the characteristics of the population of schools. In particular, such a sample can provide estimates of the variation and covariation between schools in the slope and intercept parameters and will allow us to compare schools with different characteristics.

An important class of situations arises when we wish primarily to have information about each individual school in a sample, but where we have a large number of schools so that (2.2) would involve estimating a very large number of parameters. Furthermore, some schools may have rather small numbers of students and application of (2.2) would result in imprecise and possibly widely fluctuating estimates. In such cases, if we regard the schools as members of a population and then use our population estimates of the mean and between-school variation, we can utilise this information to obtain more precise estimates for each individual school. This will be discussed later when we deal with higher level residuals.

2.2 The 2-level model

We now develop a general notation which will be used throughout this and later chapters, elaborated where necessary. We then discuss the estimation of model parameters and residuals and this is followed by illustrative examples.

To make (2.2) into a genuine 2-level model we let α_j and β_j become random variables. For consistency of notation replace α_j by β_{0j} and β_j by β_{1j} and rewrite these as

$$\beta_{0j} = \beta_0 + u_{0j}, \quad \beta_{1j} = \beta_1 + u_{1j}$$

where u_{0j}, u_{1j} are now random variables with parameters

$$\begin{aligned} E(u_{0j}) &= E(u_{1j}) = 0 \\ \text{var}(u_{0j}) &= \sigma^2_{u0}, \quad \text{var}(u_{1j}) = \sigma^2_{u1}, \quad \text{cov}(u_{0j}, u_{1j}) = \sigma_{u01} \end{aligned} \tag{2.3}$$

We can now write (2.2) in the form

$$\begin{aligned} y_{ij} &= \beta_0 + \beta_1 x_{ij} + (u_{0j} + u_{1j}x_{ij} + e_{0ij}) \\ \text{var}(e_{0ij}) &= \sigma^2_{e0} \end{aligned} \tag{2.4}$$

We shall require the extra suffix in the level 1 residual term for the models to be introduced in Chapter 3. We have expressed the response variable y_{ij} as the sum of a fixed part and a random part within the brackets and we shall also generally write the fixed part of (2.4) in the matrix form so that

$$\begin{aligned} E(Y) &= X\beta \\ \text{with} \quad Y &= \{y_{ij}\} \\ E(y_{ij}) &= X_{ij}\beta = (X\beta)_{ij}, \quad X = \{X_{ij}\} \end{aligned}$$

where β is the vector of fixed part coefficients (parameters), {} denotes a matrix, X is the design matrix for the explanatory variables and X_{ij} is the ij-th row of X. For Model (2.4) we have $X = \{1\ x_{ij}\}$. Note the alternative representation for the i-th row of the fixed part of the model.

The random variables in this model are referred to as 'residuals' and in the case of a single level model the level 1 residual e_{0ij} becomes the usual linear model residual term. The random variables are also known as 'random effects'. To make the model equation symmetrical so that each coefficient has an associated explanatory variable, we can define a further explanatory variable for the intercept β_0 and its associated residual, u_{0j}, namely x_{0ij}, a constant which takes the value 1.0. For simplicity this variable sometimes may be omitted so that the associated terms are written in the model just as β_0 and u_{0j}.

The feature of (2.4) which distinguishes it from standard linear models of the regression or analysis of variance type is the presence of more than one residual term and this implies that special procedures are required to obtain satisfactory parameter estimates. Note that it is the structure of the random part of the model which is the key factor. In the fixed part the variables can be measured at any level, for example, in the JSP data we can measure characteristics of schools or teachers. We can also include so called 'compositional' variables such as the average 8-year mathematics test score for all pupils in each school. The presence of such variables does not alter the estimation procedure, although results will require careful interpretation.

2.3 Parameter estimation

2.3.1 The variance components model

Equation (2.4) requires the estimation of two fixed coefficients, β_0, β_1, and four other parameters, σ^2_{u0}, σ^2_{u1}, σ_{u01} and σ^2_{e0}. We refer to such variances and covariances as *random parameters*. We start, however, by considering the simplest 2-level model which includes only the random parameters σ^2_{u0}, σ^2_{e0}, namely

$$y_{ij} = \beta_0 + u_{0j} + e_{0ij}$$
$$\text{var}(e_{0ij}) = \sigma^2_{e0} \quad \text{var}(u_{0j}) = \sigma^2_{u0}$$

It is termed a variance components model because the variance of the response, about the fixed component, the *fixe predictor*, is

$$\text{var}(y_{ij}|\beta_0, \beta_1, x_{ij}) = \text{var}(u_0 + e_{0ij}) = \sigma^2_{u0} + \sigma^2_{e0}$$

that is, the sum of a level 1 and a level 2 variance, where the level 1 and level 2 residuals are assumed to be mutually independent. For the JSP data this model implies that the total variance for each student is constant and that the covariance between two students (denoted by i_1, i_2) in the same school is given by

$$\text{cov}(u_{0j} + e_{0i_1 j}, u_{0j} + e_{i_2 j}) = \text{cov}(u_{0j}, u_{0j}) = \sigma^2_{u0} \tag{2.5}$$

since the level 1 residuals are assumed to be independent. The correlation between two such students is therefore

$$\rho = \frac{\sigma^2_{u0}}{(\sigma^2_{u0} + \sigma^2_{e0})}$$

which is referred to as the 'intra-level-2-unit correlation'; in this case the intra-school correlation.[1] For the variance components model this also measures the proportion of the total variance which is between-schools. In a model with three levels, say with schools, classrooms and students, we will have two such correlations and variance proportions; the intra-school correlation which is also the proportion of variance that is between schools and the intra-classroom correlation which is also that between classrooms. In more complex models with random coefficients the intra unit correlation is not equivalent to the proportion of variance at the higher level (see Chapter 3). To avoid confusion we shall use the term *variance partition coefficien* (VPC) to describe the proportion of variance at level 2 or at higher levels.

The existence of a nonzero intra-unit correlation, resulting from the presence of more than one residual term in the model, means that traditional estimation

[1] In the sample survey literature and elsewhere such as in genetics, the term 'intra-class correlation' is used, but this clearly is confusing in the educational application.

$$\begin{pmatrix} \sigma^2_{u0} + \sigma^2_{e0} & \sigma^2_{u0} & \sigma^2_{u0} \\ \sigma^2_{u0} & \sigma^2_{u0} + \sigma^2_{e0} & \sigma^2_{u0} \\ \sigma^2_{u0} & \sigma^2_{u0} & \sigma^2_{u0} + \sigma^2_{e0} \end{pmatrix}$$

Figure 2.3 Covariance matrix of three students in a single school for a variance components model.

procedures such as 'ordinary least squares' (OLS) which are used, for example, in multiple regression, are inapplicable and a later section illustrates how the application of OLS techniques can lead to incorrect inferences. We now look in more detail at the structure of a 2-level dataset, focusing on the covariance structure typified by Figure 2.3.

The matrix in Figure 2.3 is the (3 × 3) covariance matrix for the scores of three students in a single school, derived from the above expressions. For two schools, one with three students and one with two, the overall covariance matrix is shown in Figure 2.4. This 'block-diagonal' structure reflects the fact that the covariance between students in different schools is zero, and clearly extends to any number of level 2 units.

A more compact way of presenting this matrix, which we shall use again, is given in Figure 2.5, where $I_{(n)}$ is the $(n \times n)$ identity matrix and $J_{(n)}$ is the (n × n) matrix of ones. The subscript 2 for V indicates a 2-level model. In single-level OLS models σ^2_{u0} is zero and this covariance matrix then reduces to the standard form $\sigma^2 I$ where σ^2 is the (single) residual variance.

$$\begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix}$$

where

$$A = \begin{pmatrix} \sigma^2_{u0} + \sigma^2_{e0} & \sigma^2_{u0} & \sigma^2_{u0} \\ \sigma^2_{u0} & \sigma^2_{u0} + \sigma^2_{e0} & \sigma^2_{u0} \\ \sigma^2_{u0} & \sigma^2_{u0} & \sigma^2_{u0} + \sigma^2_{e0} \end{pmatrix}$$

$$B = \begin{pmatrix} \sigma^2_{u0} + \sigma^2_{e0} & \sigma^2_{u0} \\ \sigma^2_{u0} & \sigma^2_{u0} + \sigma^2_{e0} \end{pmatrix}$$

Figure 2.4 The block-diagonal covariance matrix for the response vector Y for a 2-level variance components model with two level 2 units.

$$V_2 = \begin{bmatrix} \sigma_{u0}^2 J_{(3)} + \sigma_{e0}^2 I_{(3)} & 0 \\ 0 & \sigma_{u0}^2 J_{(2)} + \sigma_{e0}^2 I_{(2)} \end{bmatrix}$$

Figure 2.5 Block-diagonal covariance matrix using general notation.

2.3.2 The general 2-level model with random coefficient

We can extend (2.4) in the standard way to include further fixed explanatory variables

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + \sum_{h=2}^{p} \beta_h x_{hij} + (u_{0j} + u_{1j} x_{1ij} + e_{0ij})$$

and more compactly as

$$y_{ij} = X_{ij}\beta + \sum_{h=0}^{1} u_{hj} z_{hij} + e_{0ij} z_{0ij} \tag{2.6}$$

where we use a new formulation for the explanatory variables in the random part of the model and write these more generally as

$$\begin{aligned} Z &= \{Z_0 Z_1\} \\ Z_0 &= \{1\} \text{ i.e. a vector of 1's} \\ \mathrm{Z}_1 &= \{x_{1ij}\} \end{aligned}$$

The explanatory variables for the random part of the model are often a subset of those in the fixed part, as here, but this is not necessary and later we shall encounter cases where this is not so. Also, any of the explanatory variables may be measured at any of the levels; for example, we may have student characteristics at level 1 or school characteristics at level 2. Examples of both are used in the data analysis in Section 2.8.

This model, with the coefficient of X_1 random at level 2, gives rise to the following typical block structure, for a level 2 block with two level 1 units. The matrix Ω_2 is the covariance matrix of the random intercept and slope at level 2. Note that we need to distinguish carefully between the covariance matrix of the responses given in Figure 2.6 and the covariance matrix of the random coefficients. We shall also refer to the intercept term as a random coefficient. The matrix Ω_1 is the covariance matrix for

the set of level 1 random coefficients; in this case there is just a single variance term at level 1. We will also write $\Omega = \{\Omega_i\}$ for the set of these covariance matrices.

$$\begin{pmatrix} A & B \\ B & C \end{pmatrix}$$

$$A = \left(\sigma^2_{u0} + 2\sigma_{u01}x_{1j} + \sigma^2_{u1}x^2_{1j} + \sigma^2_{e0}\right)$$
$$B = \left(\sigma^2_{u0} + \sigma_{u01}(x_{1j} + x_{2j}) + \sigma^2_{u1}x_{1j}x_{2j}\right)$$
$$C = \left(\sigma^2_{u0} + 2\sigma_{u01}x_{2j} + \sigma^2_{u1}x^2_{2j} + \sigma^2_{e0}\right)$$

giving

$$\begin{pmatrix} A & B \\ B & C \end{pmatrix} = X_j \Omega_2 X_j^T + \begin{pmatrix} \Omega_1 & 0 \\ 0 & \Omega_1 \end{pmatrix}$$
$$X_j = \begin{pmatrix} 1 & x_{1j} \\ 1 & x_{2j} \end{pmatrix}, \quad \Omega_2 = \begin{pmatrix} \sigma^2_{u0} & \sigma_{u01} \\ \sigma_{u01} & \sigma^2_{u1} \end{pmatrix}, \quad \Omega_1 = \sigma^2_{e0}$$

Figure 2.6 Response covariance matrix for a level 2 unit containing two level 1 units for a 2-level model with a random intercept and random regression coefficien at level 2.

We also see here the general pattern for constructing the response covariance matrix which generalises both to higher order models and, as we shall see in Chapter 3, to complex variance structures, especially those at level 1.

We now present a basic maximum likelihood (ML) procedure for obtaining estimates for our models before continuing to explore the dataset. Later in the chapter we will look at other estimation procedures.

2.4 Maximum likelihood estimation using iterative generalised least squares (IGLS)

The iterative generalised least squares (IGLS) algorithm forms the basis for many of the developments in later chapters and we now summarise the main features. Appendix 2.1 sets out the details.

Consider the simple 2-level variance components model

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + u_{0j} + e_{0ij}, \quad \text{var}(e_{0ij}) = \sigma^2_{e0}, \ \text{var}(u_{0j}) = \sigma^2_{u0} \tag{2.7}$$

Suppose that we knew the values of the variances, and so could construct immediately the block-diagonal matrix V_2, which we will refer to simply as V. We can then apply the usual Generalised Least Squares (GLS) estimation procedure to obtain the

estimator for the fixed coefficients, namely

$$\hat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} Y \tag{2.8}$$

where in this case

$$X = \begin{pmatrix} 1 & x_{11} \\ 1 & x_{21} \\ \vdots & \vdots \\ 1 & x_{n_m m} \end{pmatrix} \quad Y = \begin{pmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{n_m m} \end{pmatrix} \tag{2.9}$$

with m level 2 units and n_j level 1 units in the j-th level 2 unit. Since we are assuming that the residuals have normal distributions, (2.8) also yields maximum likelihood estimates.

Our estimation procedure is iterative. We would usually start from 'reasonable' estimates of the fixed parameters. Typically, these will be those from an initial OLS fit (that is assuming $\sigma^2_{u0} = 0$), to give the OLS estimates of the fixed coefficients $\hat{\beta}_{OLS}$. From these, we form the 'raw' residuals

$$\tilde{y}_{ij} = y_{ij} - \hat{\beta}_0^{OLS} - \hat{\beta}_1^{OLS} x_{ij} \tag{2.10}$$

and the vector of these raw residuals is written as

$$\tilde{Y} = \{\tilde{y}_{ij}\}$$

If we form the cross-product matrix $\tilde{Y}\tilde{Y}^T$ we see that, if our OLS estimates were to be unbiased (they are in fact consistent), then the expected value of this is simply V. We can rearrange this cross product matrix as a vector by stacking the columns one on top of the other which is written as $vec(\tilde{Y}\tilde{Y}^T)$ and similarly we can construct the vector $vec(V)$. For the structure given in Figure 2.4 these both have $3^2 + 2^2 = 13$ elements. The relationship between these vectors can be expressed as the following linear model

$$\begin{pmatrix} \tilde{y}^2_{11} \\ \tilde{y}_{21}\tilde{y}_{11} \\ \vdots \\ \tilde{y}^2_{22} \end{pmatrix} = \begin{pmatrix} \sigma^2_{u0} + \sigma^2_{e0} \\ \sigma^2_{u0} \\ \vdots \\ \sigma^2_{u0} + \sigma^2_{e0} \end{pmatrix} + R = \sigma^2_{u0} \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} + \sigma^2_{e0} \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 1 \end{pmatrix} + R \tag{2.11}$$

where R is a residual vector. The left hand side of (2.11) is the response vector in the linear model and the right hand side contains two explanatory variables, with coefficients $\sigma^2_{u0}, \sigma^2_{e0}$ which are to be estimated. The estimation involves an application of GLS using the estimated covariance matrix of $vec(\tilde{Y}\tilde{Y}^T)$, assuming normality, namely $2(V^{-1} \otimes V^{-1})$ where $\otimes$ is the Kronecker product. The normality assumption allows us to express this covariance matrix as a function of the random parameters. Even if the normality assumption fails to hold, the resulting estimates are still consistent, although not fully efficient, but standard errors, estimated using the normality assumption and, for example, confidence intervals will generally not be consistent. For certain variance component models alternative distributional assumptions have been studied, especially for discrete response models of the kind to be discussed in Chapter 4 (see, for example, Clayton and Kaldor, 1987) and maximum likelihood estimates obtained. For more general models, however, with several random coefficients, the assumption of multivariate normality is a flexible one which allows a convenient parameterisation for complex covariance structures at several levels. It is this assumption which forms the basis of the analyses in the remainder of this chapter, although in Section 2.13.5 we will look at a slight relaxation of this assumption where we fit the more general t-distribution for the level 1 residuals. The usefulness of the normality assumption is further extended in Chapter 7 where we study latent normal models that allow transformations of non-normally distributed variables to normality.

With the estimates obtained from applying GLS to (2.11) we return to (2.8) to obtain new estimates of the fixed effects and so alternate between the random and fixed parameter estimation until the procedure converges, that is the estimates for all the parameters do not change appreciably from one cycle to the next. At convergence we will have maximum liklelihood estimates. Essentially the same procedure is used for the more complicated models in the following chapters and is incorporated in the MLwiN program (Rasbash *et al*., 2009).

The maximum likelihood procedure in fact produces biased estimates of the random parameters because it takes no account of the sampling variation of the fixed parameters. This may be important in small samples, and we can produce unbiased estimates by using a modification known as restricted maximum likelihood (REML). The IGLS algorithm is readily modified to produce these restricted (RIGLS) estimates (Appendix 2.1).

Longford (1987) developed a procedure based upon a 'Fisher scoring' algorithm and Raudenbush (1994) shows that this is formally equivalent to IGLS. A variation on IGLS is expected generalised least squares (EGLS). This focuses interest on the fixed part parameters and uses the estimate of V obtained after the first iteration merely to obtain a consistent estimator of the fixed part coefficients without further iterations. A variant of this separates the level 1 variance from V as a parameter to be estimated iteratively along with the fixed part coefficients.

Another algorithm for obtaining ML or REML estimates is the EM algorithm or variants of it (Bryk and Raudenbush, 2002). This is outlined in Appendix 2.3 and has been incorporated into several software packages, partly because of its computational simplicity. In Chapter 4 we shall look at maximum likelihood and quasilikelihood procedures for generalised linear models with discrete responses.

2.5 Marginal models and generalised estimating equations (GEE)

It is worth emphasising the distinction between multilevel models, sometimes referred to in this context as 'subject specific' models, and so-called 'marginal' models such as the GEE model (Zeger *et al*., 1988; Liang *et al*., 1992). When dealing with hierarchical data these latter models typically start with a formulation for the covariance structure, for example, but not necessarily, based upon a multilevel structure, and aim to provide estimates with acceptable properties only for the fixed parameters in the model, treating the existence of any random parameters as a necessary 'nuisance'. More specifically, the estimation procedures used in marginal models are known to have useful asymptotic properties in the case where the exact form of the random structure is unknown.

If interest lies only in the fixed parameters, marginal models can be useful since they give directly unbiased estimates for these parameters. Even here, however, they may be inefficient if they utilise a covariance structure that is substantially incorrect. They are, however, generally more robust than multilevel models to serious misspecification of the covariance structure (Heagerty and Zeger, 2000). Fundamentally, however, marginal models address different research questions. From a multilevel perspective, the failure explicitly to model the covariance structure of complex data is to ignore information about variability that, potentially, is as important as knowledge of the average or fixed effects. Thus, in a repeated measures growth study, knowledge of how individual growth rates vary, possibly differentially according to say demographic factors, will be important information and in Chapter 5 we will show how such information can be used to provide efficient predictions in the case of human growth.

Also, when we discuss multilevel models for discrete response data in Chapter 4 we will show how to obtain estimates for population or subpopulation means equivalent to those obtained from marginal models. In the case of normal response linear multilevel models, GEE and multilevel models lead to the same fixed coefficient estimates. For a further discussion of the limitations of marginal models, see the paper by Lindsey and Lambert (1998).

2.6 Residuals

In a single level model such as (2.1), the usual estimate of the single residual term e_i is just $\tilde{y}_i$ the raw residual. In a multilevel model, however, we shall generally have several residuals at different levels. We can consider estimates for the individual residuals along the following lines.

Given the parameter estimates, consider predicting a specific residual, say u_{0j} in a 2-level variance components model. Specifically we require for each level 2 unit

$$\hat{u}_{0j} = E(u_{0j}|Y, \hat{\beta}, \hat{\Omega}) \tag{2.12}$$

We shall refer to these as estimated or predicted residuals or, using Bayesian terminology, as conditional posterior residual estimates. If we ignore the sampling variation attached to the parameter estimates in (2.12) we have

$$\begin{aligned} \operatorname{cov}(\tilde{y}_{ij}, u_{0j}) &= \operatorname{var}(u_{0j}) = \sigma^2_{u0} \\ \operatorname{cov}(\tilde{y}_{ij}, e_{0ij}) &= \sigma^2_{e0} \\ \operatorname{var}(\tilde{y}_{ij}) &= \sigma^2_{u0} + \sigma^2_{e0} \end{aligned} \tag{2.13}$$

We may regard (2.12) as a (linear) regression of u_{0j} on the set of raw residuals $\{\tilde{y}_{ij}\}$ for the j-th level 2 unit and (2.13) defines the quantities required to estimate the regression coefficients and hence $\hat{u}_{0j}$. Full details are given in Appendix 2.2. For the variance components model we obtain

$$\begin{aligned} \hat{u}_{0j} &= \frac{n_j \sigma^2_u}{(n_j \sigma^2_u + \sigma^2_{e0})} \tilde{y}_j \\ \tilde{e}_{0ij} &= \tilde{y}_{ij} - \hat{u}_{0j} \\ \tilde{y}_j &= (\sum_i \tilde{y}_{ij})/n_j \end{aligned} \tag{2.14}$$

where n_j is the number of level 1 units in the j-th level 2 unit. Estimates are obtained by substituting sample values in (2.14). The factor multiplying the mean ($\tilde{y}_j$) of the raw residuals for the j-th unit is often referred to as a 'shrinkage factor' since it is always less than or equal to one in absolute value. As n_j increases this factor tends to one, and as the number of level 1 units in a level 2 unit decreases the 'shrinkage estimator' of u_{0j} becomes closer to zero. In many applications the higher level residuals are of interest in their own right and the increased shrinkage for a small level 2 unit can be regarded as expressing the relative lack of information in the unit so that the best estimate places the predicted residual close to the overall population value as given by the fixed part.

These residuals therefore can have two roles. Their basic interpretation is as random variables with a distribution whose parameter values tell us about the variation among the level 2 units, and provide efficient estimates for the fixed coefficients. A second interpretation is as individual estimates for each level 2 unit where we use the assumption that they belong to a population of units with a known (estimated) distribution to predict their values. In particular, for units which have only a few level 1 units, we can obtain more precise estimates than if we were to ignore the population membership assumption and use only the information from those units. This becomes especially important for estimates of residuals for random coefficients, other than the intercept, where, in the extreme case of only one level 1 unit in a level 2 unit, we lack information to form an independent estimate. We illustrate this when we consider predictions based upon repeated measures growth models.

As in single level models, we can use the estimated residuals to help check on the assumptions of the model. The two particular assumptions that can be studied readily are the assumption of normality and that the variances in the model are constant

across values of the explanatory variables. It is common to standardise the residuals by dividing by the appropriate standard errors, which are functions of the explanatory variables for the random effects and the sizes of the units; the formulae for these are given in Appendix 2.2, where we refer to them as 'diagnostic' standard errors.

When the residuals at higher levels are of interest in their own right, we need to be able to provide interval estimates and significance tests as well as point estimates for them or functions of them. For these purposes we require estimates of the standard errors of the estimated residuals, where the sample estimate is viewed as a random realisation from repeated sampling of the higher level units whose unknown true values are of interest. The formulae for these 'conditional' or 'comparative' standard errors are also given in Appendix 2.2. For the estimates given in 2.12 the comparative variance is given by $\sigma_e^2\sigma_u^2(\sigma_e^2 + n_j\sigma_u^2)^{-1}$ and the diagnostic variance by $n_j\sigma_u^4(\sigma_e^2 + n_j\sigma_u^2)^{-1}$.

The level 1 residuals are generally not of interest in their own right but are used rather for model checking, having first been standardised using the diagnostic standard errors.

2.7 The adequacy of ordinary least squares estimates

When a variance partition coefficient is small, we can expect reasonably good agreement between the multilevel estimates and the simpler OLS ones. While it is difficult to give general guidelines about when OLS is an adequate alternative, we can readily derive an explicit formula for the balanced 2-level variance component model using a simple regression equation with an intercept and a single explanatory variable

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + u_j + e_{ij}$$

Write ρ_y ρ_x for the intra-unit correlations for Y, X respectively and n for the number of level 1 units in each level 2 unit. To obtain an estimate of the correct standard error for the estimate of β_1, we multiply the usual OLS estimate of the standard error by the quantity

$$\{1 + \rho_y\rho_x(n-1)\}^{1/2}$$

Thus, if there is exactly one level 1 unit per level 2 unit or either of the intra-unit correlations are zero, this expression is equal to 1.0 and the usual expression is correct. As n increases so the OLS estimator increasingly underestimates the true standard error. Thus, with $\rho_y = \rho_x = 0.20$ and 76 level 1 units per level 2 unit, the true standard error is, on average, twice the OLS estimate. Hence, confidence intervals based on the OLS estimate will be too short and significance tests will too often reject the null hypothesis. By designing a study where n is small, we may be able to rely on OLS procedures to give adequate estimates for the fixed coefficients, but this would then not allow us to study any multilevel structures with adequate precision. Hedges (2009) derives correction factors for testing significance and constructing confidence

Table 2.1 Variance components model applied to JSP data.

Parameter	Estimate (s.e.)	OLS Estimate (s.e.)
Fixed		
Intercept	13.9	13.8
Age 8 score	0.65 (0.025)	0.65 (0.026)
Random		
σ^2_{u0} (between-schools)	3.19 (1.0)	
σ^2_{e0} (between-students)	19.8 (1.1)	23.3 (1.2)
Variance partition coefficient	0.14	
Deviance	4294.2	4357.3

intervals in 3-level models, but these are of limited utility since they apply only for variance component models.

2.8 A 2-level example using longitudinal educational achievement data

We shall fit the simple 2-level variance components model (2.7) to the JSP data with the maths score at age 11 as the response and a single explanatory variable, the maths score at age 8, in addition to the constant term, equal to 1 and defining the intercept. The parameter values are displayed in Table 2.1 with the ordinary least squares estimates given for comparison.

Comparing the OLS with the multilevel estimates, we see that the fixed coefficients are similar, but that there is a variance partition coefficient value of 0.14. The estimate of the standard error of the between school variance is less than a third of the variance estimate, suggesting a value highly significantly different from zero[2]. This comparison, however, should be treated cautiously, since the variance estimate does not have a normal distribution and the standard error is only estimated, although the size of the sample here will make the latter caveat less important. It is generally preferable to carry out a likelihood ratio test by estimating the 'deviance' for the current model and the model omitting the level 2 variance (see McCullagh and Nelder, 1989) and the next section will deal more generally with inference procedures. The difference between the deviances is 63.1. This value would normally be referred to tables of the chi-squared distribution with one degree of freedom, and is highly significant. In the present case, however, the null hypothesis of a zero variance is on the 'boundary' of the feasible parameter space; we do not envisage a negative variance. In this case, the P-value to be used is half the one obtained from the tables

[2] Since the intercept estimate depends upon the origins chosen for the other explanatory variables, its value generally has no substantive interest and we omit its standard error.

of the chi-squared distribution (Shapiro, 1985). Note that if we use the standard error estimate given in Table 2.1 to judge significance we obtain the corresponding value of $(3.19/1.0)^2 = 10.2$ which, in this case, is very much smaller than the likelihood ratio test statistic. Note also that if we use this test we would again use half the nominal P-value since only positive departures are possible.

A similar issue arises when simultaneously testing several parameters where one or more is constrained in this way, such as a variance plus covariances. In this case, the appropriate test statistic is a weighted mixture of chi-squared distributions and details are given by Shapiro (1985). In the common case where we are testing a set of covariances and a single associated variance, this reduces to the following.

Suppose that we have r covariance parameters and adopt a 5 % significance level. Then the critical value, c, for judging significance is given by the following formula for a mixture of two chi-squared distributions

$$0.5 \times [pr(\chi^2_r \geq c) + pr(\chi^2_{r+1} \geq c)] = 0.05$$

Thus, if $r = 2$ we find that $c = 7.0$ which compares with the critical value using the conventional test with 3 degrees of freedom, of 7.8, and in general we will more often judge significance than using the conventional test. This modified chi-squared test is also known as the chi-bar test statistic and when we use it we shall quote the degrees of freedom as $r + 1$.

We now elaborate the model by adding two more explanatory variables, gender and social class. The results are set out in the first column of Table 2.2.

Here the random parameter estimates are hardly changed, nor is the coefficient of the maths score at age 8. The gender difference is very small and in favour of the girls, but is far from the conventional 5 % significance level. The social class difference favours the children of parents with non-manual occupations. When we are judging the fixed effects, a simple comparison of the estimate with its standard

Table 2.2 Variance components model applied to JSP data with gender and social class.

Parameter	Estimate (s.e.)	Estimate (s.e.)
Fixed		
Intercept	14.9	32.9
Age 8 score	0.64 (0.025)	
Gender (boys-girls)	−0.36 (0.34)	−0.39 (0.47)
Social Class (non-manual-manual)	0.72 (0.39)	2.93 (0.51)
Random		
σ^2_{u0} (between-schools)	3.21 (1.0)	4.52 (1.5)
σ^2_{e0} (between-students)	19.6 (1.1)	37.2 (2.0)
Variance partition coefficient	0.14	0.11

error is usually adequate. Because the model adjusts for the earlier maths score we can interpret the social class and gender differences in terms of the relative *progress* of girls versus boys or non-manual versus manual children. The second column in Table 2.2 shows the effects when age 8 maths score is removed from the model and the interpretation is now in terms of the actual differences found at age 11. Note that the level 1 and level 2 variances are increased, reflecting the importance of the earlier score as a predictor, and the variance partition coefficient is slightly reduced. The social class difference is much larger, suggesting that most of the difference is that existing at age 8 with a somewhat greater progress made between ages 8 and 11 years by those in the non-manual group. The gender difference remains small.

The age 8 score has been used as it stands, without centring it in any way. This is acceptable in the present case, although the strict interpretation of the intercept is the predicted score at age 11 of a child with an age 8 score of zero, which is outside the range of the observed values. If we were to measure the 8-year-score about its mean, the intercept would then be interpreted as the predicted value at the mean age 8 score in the sample. When we introduce random coefficients we shall see that this becomes an important consideration.

2.8.1 Checking for outlying units

Figure 2.7 is a plot of the estimated residuals against equivalent normal scores (see section 2.11) and shows one school, identified as number 38, with the largest residual of 3.5. It is often useful to study the effect of omitting one or more units from an

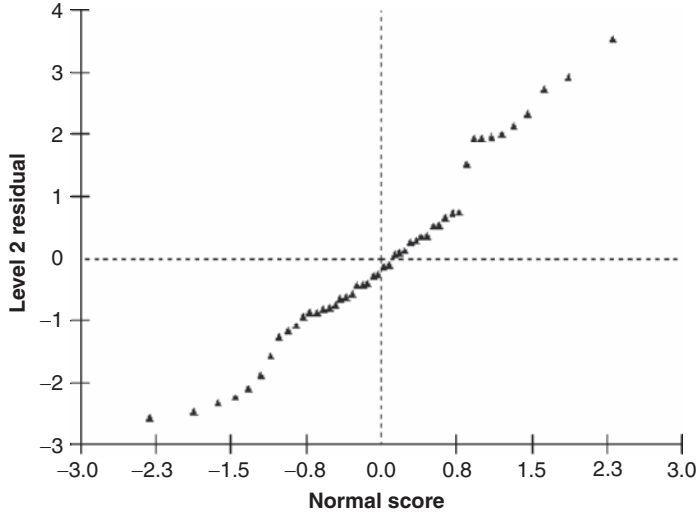


Figure 2.7 Standardised level 2 residuals by normal equivalent scores for first column model in Table 2.2

Table 2.3 As Table 2.2, Model A omitting school 38; Model B fitting a constant for school 38.

Parameter	Estimate (s.e.) Model A	Estimate (s.e.) Model B
Fixed		
Intercept	14.5	14.7
Age 8 score	0.65 (0.026)	0.64 (0.025)
Gender (boys-girls)	−0.40 (0.34)	−0.37 (0.34)
Social Class (non-manual-manual)	0.74 (0.39)	0.72 (0.38)
School 38		6.1 (1.5)
Random		
σ^2_{u0} (between-schools)	2.74 (0.9)	2.75 (0.9)
σ^2_{e0} (between-students)	19.6 (1.1)	19.6 (1.1)
Variance partition coefficient	0.12	0.12

analysis to see what difference this makes to the parameter estimates and we shall say more about this below. For now we illustrate the effect of omitting such units using school 38. Table 2.3 shows the parameter estimates associated with two different procedures.

In Model A, school 38 is simply omitted. The principal effect is to reduce the level 2 variance by about 14 %, with little effect on the other parameters. In Model B, we have retained all the data in the analysis, but removed school 38 from the level 2 variation by fitting a separate Intercept term in the fixed part of the model. For the explanatory variable defining the level 2 variance we fit Z_0^* rather than Z_0, where

$$Z_0^* = \left\{ \begin{array}{l} 0 \text{ if school } 38 \\ 1 \text{ otherwise} \end{array} \right\}$$

The relatively small number of students, nine, in school 38 accounts for the fact that its shrunken residual mean of 3.5 is considerably less than the directly fitted mean of 6.1. Although it makes little difference to the parameter estimates in this example, in general it seems preferable to fit separate parameters for influential units and retain as much data as possible in the analysis.

2.8.2 Model checking using estimated residuals

We now check a particularly important assumption of the model by looking at the residuals. Figure 2.8 is a plot of the standardised level 1 residuals against the fixed part predicted value and Figure 2.9 is a plot of these residuals against their equivalent normal scores. The latter plot is close enough to a straight line to give us some confidence that the normality assumption is reasonable. Figure 2.8, however, shows

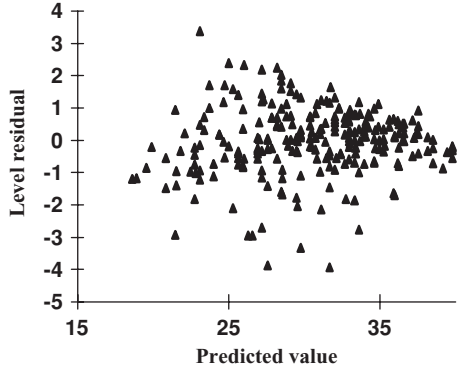


Figure 2.8 Standardised level 1 residuals by predicted values for Table 2.2.

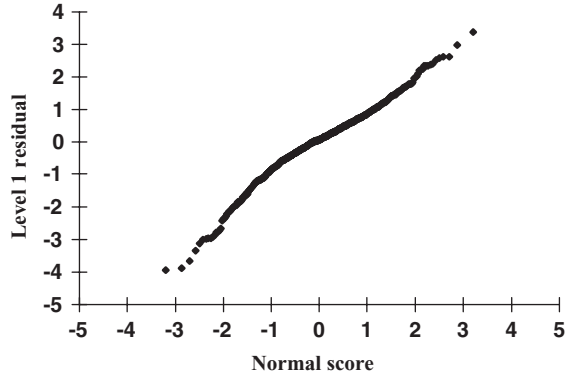


Figure 2.9 Standardised level 1 residuals by normal equivalent scores for Table 2.2

the same pattern as Figure 2.1 of a decreasing variance with increasing 8-year score, so that the assumption of a constant level 1 variance is clearly untenable.

In Chapter 3, we look at ways of directly modelling such nonconstant or complex level 1 variation. In Section 2.9, we look at general diagnostic procedures and in Section 2.10 at nonlinear transformations of the response variable to produce a more constant variance.

2.9 General model diagnostics

Procedures for model exploration fall into two broad groups. The first of these is concerned with choosing the distributional assumptions, the set of variables for inclusion

and the scales upon which they are measured. The choice of variables for inclusion, where external subject matter considerations are unavailable, is more complicated in multilevel models than in single level models. In single level linear models, stepwise procedures are available, which if used with care can produce usefully descriptive models. In multilevel models we need to consider random parameters as well as fixed coefficients and efficient procedures for the partial automation of stepwise techniques do not seem to be available.

The second group of procedures is based upon analyses of the effects that individual units can have on parameter estimates. We have already explored some uses of the estimated residuals and how the omission or modification of units can change inferences. A related approach uses measures of influence, and is described in detail by Langford and Lewis (1998) and implemented in the MlwiN package (Rasbash *et al*., 2009). A somewhat different approach to model diagnosis is taken by Hodges (1998) who considers in particular the use of MCMC chains.

The notion of 'influence' is that of introducing small changes to a model and assessing the effects of these on the resulting inferences. A central feature is 'leverage' which expresses the effect of deleting a particular unit on the parameter estimates. In a 2-level model Langford and Lewis use the set of values for the level 1 units in level 2 unit j given by

$$H_j = diag(V_j^{-0.5} X_j (X_j^T V_j^{-1} X_j)^{-1} X_j^T V_j^{-0.5})$$

for the fixed parameters and an analogous formula for the random parameters. The larger the values the more potential they have for affecting the model parameter estimates.

Langford and Lewis also consider 'deletion' residuals, formed by deleting a particular unit, estimating from the reduced model and applying these parameter estimates to the complete dataset to calculate the residual for the deleted unit. This is approximately equivalent to fitting a dummy variable for the unit in question. They also discuss leverage values associated with random coefficients which may help the data analyst to assess the effect of a change in one or more of these. A detailed example explaining the use of these procedures is given by Lewis and Langford (2001) and Shi and Chen (2008) discuss unit deletion methods.

When using diagnostic procedures some care is needed. Choosing to omit a unit or fit a dummy variable for it may perhaps best be regarded as part of a sensitivity analysis to see how robust the model is to varying the model assumptions. The use of formal significance tests when units have been detected on the basis that they are extreme also needs to be done carefully. A simple 'conservative' procedure that can be used when identifying outlying units is to multiply the significance levels (P values) obtained for each test by the number of units at that level. Thus, from Table 2.3 the significance level associated with fitting a separate intercept for school 38 is 0.000023, and when multiplied by the number of schools (48) is 0.0011, which is still highly significant.

2.10 Higher level explanatory variables and compositional effects

We have already mentioned that from the point of view of estimating parameters, the explanatory variables can be defined or measured at any level. For substantive interpretations, however, explanatory variables measured at levels 2 or above often have particular interpretations. We illustrate some of these using the JSP dataset and forming the explanatory variable which is the mean age 8 maths score. This is often known as a 'compositional' variable since it measures an aspect of the composition of the school to which the individual student belongs. We are interested in whether the average age 8 score has an effect on the age 11 score, after having adjusted for the student's own age 8 score. For this analysis all the age 8 scores are measured about the sample mean value of 25.98 (see Table 2.4). Model A adds the average school age 8 score. Its coefficient is very small and not significant. Model B uses the school centred age 8 score. This is often advocated on the grounds that it is the difference between a student's score and the average score for that student's school which is likely to be the most relevant predictor of later achievement.

Bryk and Raudenbush (2002) give a detailed discussion of this issue for models where the compositional variable, as here, is a mean computed for all the students in the school, or more generally all the level 1 units in the relevant level two unit. Models A and B are, of course, formally equivalent and Model A indicates directly

Table 2.4 Variance components model for JSP data with mean age 8 score measured about sample mean and centring about school mean.

Parameter	Estimate (s.e.)	Estimate (s.e.)	Estimate (s.e.)
	Model A	Model B	Model C
Fixed			
Intercept	31.5	31.5	31.7
Age 8 score	0.64 (0.025)		1.25 (0.26)
Age 8 score centred on school mean		0.64 (0.026)	
Gender (boys-girls)	−0.36 (0.34)	−0.36 (0.34)	−0.37 (0.34)
Social Class (non-manual-manual)	0.72 (0.38)	0.72 (0.31)	0.79 (0.31)
School mean age 8 score	−0.01 (0.13)	0.63 (0.12)	−0.03 (0.12)
Age 8 score × school mean age 8 score			−0.02 (0.01)
Random			
σ^2_{u0} (between-schools)	3.21 (1.0)	3.21 (1.0)	3.13 (1.0)
σ^2_{e0} (between-students)	19.6 (1.1)	19.6 (1.0)	19.5 (1.1)
Variance partition coefficient	0.14	0.14	0.14

that a simpler model omitting the school mean score is adequate. It is Model C, as discussed below, which introduces a more complex model.

In fact, the mean score for students in a school is only one particular summary statistic describing the composition of the students. Another summary would be the spread of scores, measured, for example, by their standard deviation. We can also consider measures such as the proportions of high or low scoring students and in general any set of such measures. When using the average score we can also consider using the median or modal score rather than the mean. With any of these other measures we may wish to retain the deviation from the school mean as an explanatory variable, and we could even consider introducing a more complex function of this, for example, by adding higher order terms. This is a fruitful area for data analysis.

Model C looks at the possibility of an interaction between student score and school mean and we do find a significant effect which we can interpret as follows. The higher the school mean age 8 score the lower the coefficient of the student's age 8 score. One implication of this is that for two relatively low scoring students at age 8 years, the one in the school with a higher average is predicted to do better at age 11 years. To study this further we now need to introduce a model with random coefficients where we explicitly allow each school's coefficient to vary randomly at level 2, as in (2.6); see Table 2.5.

The addition of the age 8 score coefficient as a random variable at level 2 somewhat increases the social class difference and somewhat decreases the gender difference, but within their standard errors.

The level 1 variance is reduced and we have significant 'slope' variation at level 2; the likelihood ratio test criterion is 52.4 which is well beyond the adjusted 5 %

Table 2.5 Random coefficient model for JSP data.

Parameter	Estimate (s.e.)
Fixed	
Intercept	31.7
Age 8 score	1.11 (0.35)
Gender (boys-girls)	−0.25 (0.32)
Social Class (non-manual-manual)	0.96 (0.36)
School mean age 8 score	−0.04 (0.13)
Age 8 score × school mean age 8 score	−0.02 (0.01)
Random	
Level 2	
σ^2_{u0} (Intercept)	3.67 (1.03)
σ_{u01} (covariance)	−0.34 (0.09)
σ^2_{u1} (age 8 score)	0.03 (0.01)
Level 1	
σ^2_{e0}	17.7 (1.0)

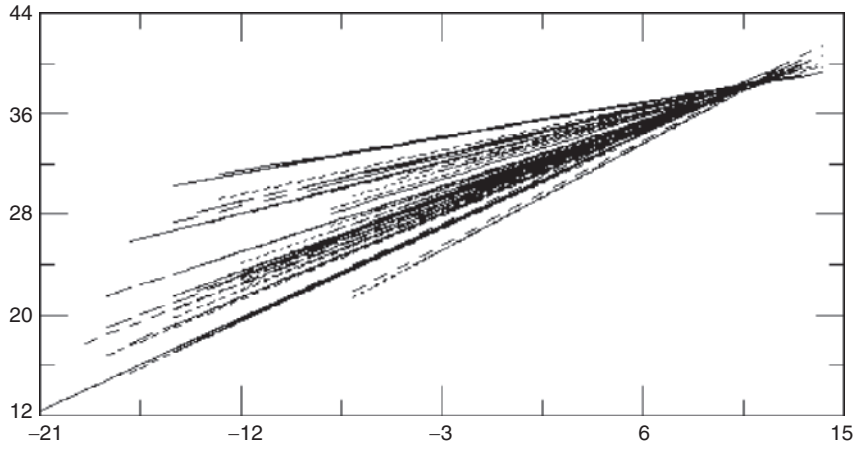


Figure 2.10 Predicted 11-year score by 8-year score for JSP schools.

level of 7.0 as discussed in Section 2.8. If we calculate the correlation between the intercept and slope at level 2 we obtain a value of –1.03! This can happen as a result of sampling variation and implies that the population correlation is very high; the basic IGLS algorithm does not impose an explicit constraint on this correlation. In fact, we shall see in Chapter 3 that we can constrain this correlation to be exactly –1.0 and thus admissible. Alternatively, by suitably elaborating the model or by carrying out certain transformations we can avoid this problem. Also, if we are using MCMC estimation (see Section 2.13) this problem will not arise. For now, however, in order to illustrate what such a high correlation means in the present data we can compute residuals for each school, for the slope and intercept. With these estimates we can then predict the age 11 score for any set of values of the explanatory variables. Figure 2.10 shows the predicted values for manual girls by age 8 score.

The predicted lines for the high scores at 8 years of age are very close together, separating as the age 8 score decreases. The slope residual is almost uncorrelated (–0.02) with the mean age 8 score and the compositional coefficient of mean age 8 score is little changed. We can add, therefore, to the previous compositional effect, the statement that some schools are differentially 'effective' for pupils with low age 8 scores, with little difference for high age 8 scores. We continue to analyse this dataset in Chapter 3, showing how further elaboration of the variance structure of the model leads to certain simplifications of interpretation.

2.11 Transforming to normality

For single level data there is a considerable literature on ways of transforming measurements to satisfy the standard model assumption of normality with constant variance (see Box and Cox, 1964). Cole and Green (1992), for example, describe a procedure for smoothly modelling the mean, variance and skewness of

measurements repeated over time and using the estimated smoothed relationships to make the appropriate transformation to normality. We shall use a general purpose procedure designed to produce approximately normally distributed residuals and study its effect on our example analyses. We discuss a very general set of procedures for transforming to normality in Chapter 7.

If we look back at Figure 2.9, we can imagine 'stretching' the Y axis to make the plot follow a straight line. Of course, we cannot do this directly since the residuals are estimates rather than observed measurements, but we can do this for the original response measurements with the expectation that this will give us residuals which are more nearly normally distributed, and also with a more constant variance. To do this, we rank all the response measurements and assign to each one the equivalent value from a standard normal distribution that cuts off the same proportion of the population as does the ranked observation. Thus, for example, if we have 99 measurements, the smallest value is a nonparametric estimate of the first percentile of the population distribution; the value such that just 1 % lie below this value. We would then assign this observation the equivalent normal score of –2.33 which is the lower 1 % point of the $N(0, 1)$ distribution. This procedure will often be justified, for example, with educational test scores, where the measurement scales themselves are essentially arbitrary, but in other cases, for example, with physical measurements, the original scale is meaningful and in such cases, we may prefer to use the variance modelling methods described in Chapter 3 or the procedures in Chapter 7.

We now repeat the analysis in Table 2.2 for our transformed data, where we have also transformed the age 8 scores in similar fashion. Table 2.6 shows the results.

The inferences we would make from this table are similar to the earlier ones, except that the social class coefficient is now more significant and the intra-school correlation has increased slightly. Figure 2.11, however, shows that the variance of the level 1 residuals is now more nearly constant, although not entirely so, and Figure 2.12 shows that the residuals follow a normal distribution more closely also.

Table 2.6 Variance components model applied to JSP data with gender and social class and normalised 11-year and 8-year scores.

Parameter	Estimate (s.e.)
Fixed	
Intercept	0.129
Age 8 score	0.668 (0.026)
Gender (boys-girls)	−0.048 (0.050)
Social Class (non-manual-manual)	0.138 (0.057)
Random	
σ^2_{u0} (between-schools)	0.080 (0.023)
σ^2_{e0} (between-students)	0.422 (0.023)
Variance partition coefficient	0.16

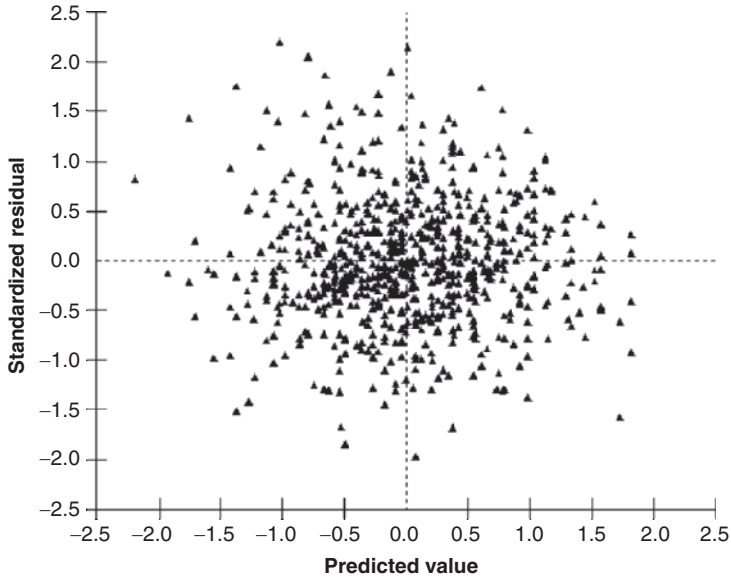


Figure 2.11 Standardised level 1 residuals by predicted value for Table 2.6.

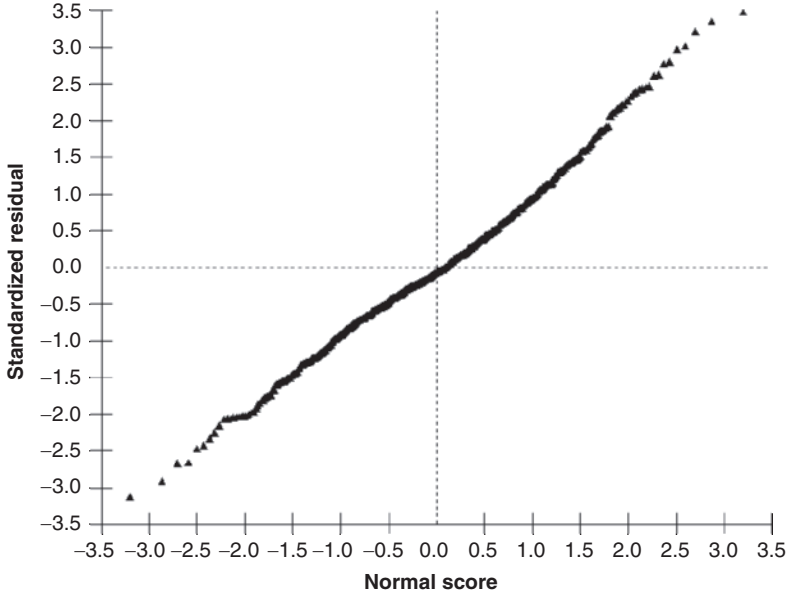


Figure 2.12 Standardised level 1 residuals by normal equivalent scores for Table 2.6.

2.12 Hypothesis testing and confidenc intervals

In this section we deal with large sample procedures for constructing interval estimates for parameters or linear functions of parameters and for hypothesis testing. Hypothesis tests are used sparingly throughout this book, since the usual form of a null hypothesis, that a parameter value or a function of parameter values is zero, is usually intrinsically implausible and also relatively uninteresting. With large enough samples a null hypothesis will almost certainly be rejected. The exception to this is where we are interested in whether a difference is positive or negative, and this is discussed in the section on residuals below. Confidence intervals emphasise the uncertainty surrounding the parameter estimates and generally provide more useful summaries for substantive interpretations.

2.12.1 Fixed parameters

In the analyses of Section 2.11 we presented parameter estimates for the fixed part parameters together with their standard errors. These are adequate for hypothesis testing or confidence interval construction separately for each parameter. In many cases, however, we are interested in combinations of parameters. For hypothesis testing, this most often arises for grouped or categorised explanatory variables where n group effects are defined in terms of $n-1$ dummy variable contrasts and we wish simultaneously to test whether these contrasts are zero. In the case of the analysis in Table 2.2 we may be interested in the hypothesis that the gender and social class effects taken jointly, are zero. We may also be interested in providing a pair of confidence intervals for the parameter estimates. We proceed as follows.

Define a *(r x p)* contrast matrix $\boldsymbol{C}$. This is used to form linearly independent functions of the p fixed parameters in the model of the form $f = C\beta$, so that each row of $\boldsymbol{C}$ defines a particular linear function. Parameters which are not involved have the corresponding elements set to zero. Suppose we wish to test the hypothesis in Table 2.2 that the gender and social class coefficients are jointly zero. We define

$$C = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad f = \begin{pmatrix} \beta_2 \\ \beta_3 \end{pmatrix}$$

and the general null hypothesis is

$$H_0 : f = k, \quad k = \{0\} \text{ here}$$

We form

$$R = (\hat{f} - k)^T [C(X^T \hat{V}^{-1} X)^{-1} C^T]^{-1} (\hat{f} - k) \tag{2.15}$$

where

$$\hat{f} = C\hat{\beta}$$

If the null hypothesis is true R is distributed as approximately χ^2 with r degrees of freedom. This is often known as a 'Wald test'. Note that the term $(X^T \hat{V}^{-1} X)^{-1}$ is the estimated covariance matrix of the fixed coefficients.

If we find a statistically significant result we may wish to explore which particular linear combinations of the coefficients involved are significantly different from zero. The common instance of this is where we find that n groups differ and we wish to carry out all possible pairwise comparisons. A simultaneous comparisons procedure which maintains the overall type I error at the specified level involves carrying out the above procedure with either a subset of the rows of C or a set of (less than r) linearly independent contrasts. The value of R obtained is then judged against the critical values of the chi-squared distribution with r degrees of freedom.

We can also obtain a $(100 - \alpha)\%$ confidence region for the parameters by setting $\hat{R}$ equal to the $\alpha\%$ tail region of the χ^2 distribution with r degrees of freedom in the expression

$$\hat{R} = (f - \hat{f})^T [C(X^T \hat{V}^{-1} X)^{-1} C^T]^{-1} (f - \hat{f})$$

This yields a quadratic function of the estimated coefficients, giving an r-dimensional ellipsoidal region. For Table 2.2 we obtain the following results.

The null hypothesis test gives a value for chi-squared on 2 degrees of freedom of 4.51 with a corresponding P-value of 0.10. The 95 % confidence region is the ellipse

$$8.3(\beta_1 + 0.36)^2 + 0.22(\beta_1 + 0.36)(\beta_2 - 0.72) + 6.7(\beta_2 - 0.72)^2 = 5.99$$

where the subscripts (1,2) refer to gender and social class respectively and 5.99 is the 5 % point of the χ^2_2 distribution. Figure 2.13 displays this region, which contains the point (0,0) so that the null hypothesis that $\beta_1 = \beta_2 = 0$ is not rejected at the 5 % level.

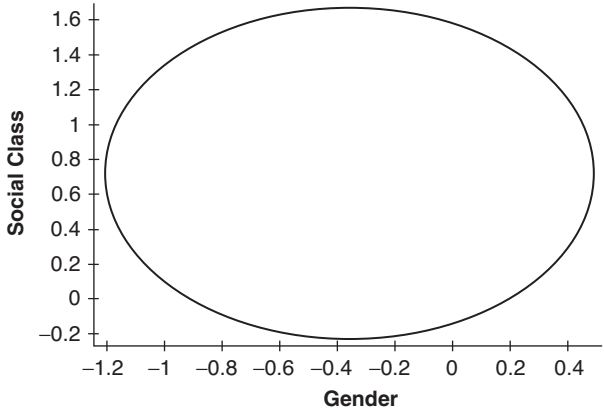


Figure 2.13 95 % confidenc region for coefficient of Social Class and Gender.

In some situations we may be interested in separate confidence intervals for all possible linear functions involving a subset of $\boldsymbol{q}$ parameters or $\boldsymbol{q}$ linearly independent functions of the parameters, while maintaining a fixed probability that all the intervals include the population value of these functions of the parameters. As before, this may arise when we have an explanatory variable with several categories and we are interested in intervals for sets of contrasts. For a $(100-\alpha)\%$ interval write C_i for the i-th row of C, then a simultaneous $(100-\alpha)\%$ interval for $C_i\beta$, for all C_i is given by

$$(C_i\hat{\beta} - d_i, C_i\hat{\beta} + d_i)$$

where

$$d_i = [C_i(X^T\hat{V}^{-1}X)^{-1}C_i^T\chi^2_{q,(\alpha)}]^{0.5}$$

where $\chi^2_{q,(\alpha)}$ is the $\alpha\%$ point of the χ^2_q distribution. For the model in the first column of Table 2.2 we obtain the following 95 % intervals for the coefficients of gender and social class, first the separate intervals then the simultaneous ones which are some 25 % wider.

$$\begin{pmatrix} -0.36 \pm 0.66 \\ 0.72 \pm 0.76 \end{pmatrix}, \quad \begin{pmatrix} -0.36 \pm 0.83 \\ 0.72 \pm 0.94 \end{pmatrix}$$

We can also use the likelihood ratio test criterion for testing hypotheses about the fixed parameters, although generally the results will be similar. The difference arises because the random parameter estimates used in (2.15) are those obtained for the full model rather than those under the null hypothesis assumption, although this modification can easily be made. For example the likelihood ratio test for gender and social class yields a value of 5.5 compared with the above value of 4.5. We shall discuss the likelihood ratio test in the next section dealing with the random parameters.

2.12.2 Random parameters

In very large samples, it is possible to use the same procedures for hypothesis testing and confidence intervals as for the fixed parameters. Generally, however, procedures based upon the likelihood statistic are preferable. To test a null hypothesis H_0 against an alternative H_1, involving the fitting of additional parameters, we form the log likelihood ratio or deviance statistic

$$D_{01} = -2\log_e(\lambda_0/\lambda_1) \tag{2.16}$$

where λ_0, λ_1 are the likelihoods for the null and alternative hypotheses and this is then referred to tables of the chi-squared (χ^2) distribution with degrees of freedom equal to the difference (q) in the number of parameters fitted under the two models. We have already quoted this statistic for testing the level 2 variance in Table 2.1, with the caveat needed when the null model involves a 'boundary' value for the parameters

(Section 2.8) and the chi-bar test is used. A similar caveat applies to the calculation of confidence regions.

We can also use (2.16) as the basis for constructing a $(100 - \alpha)\%$ confidence region for the additional parameters. If D_{01} is set to the value of the $\alpha\%$ point of the chi-squared distribution with q degrees of freedom, then a region is constructed to satisfy (2.16), using a suitable search procedure. This is a computationally intensive task, however, since all the parameter estimates are recomputed for each search point.

An alternative is to use the 'profile likelihood' (McCullagh and Nelder, 1989). In this case the likelihood is computed for a suitable region containing values of the random parameters of interest, for fixed values of the remaining random parameters.

Unlike ML estimates, we cannot use REML (RIGLS) estimates to compute a likelihood that can then be used in a general way to compare models, although the REML likelihood *can* be used if we are just making comparisons between models that have the same set of fixed effects but different random parameters. Thus, for model exploration and hypothesis testing it is recommended that ML (IGLS) estimation should be used, with a final REML estimation if this is required. When we look at MCMC estimation in Section 2.13, we see how exact interval estimates can be constructed; and in Chapter 3, we see how this can be done using the bootstrap.

2.12.3 Hypothesis testing for non-nested models

So far we have considered whether we should accept a sub-model from a model with extra parameters, for example, whether a variance term is significantly different from zero. In some cases, however, we may wish to compare two models where one is not a strict subset of the other; such models are referred to as 'non-nested'. Consider the simple variance components model for the JSP data from Table 2.1 where we use transformations of the 11-year and 8-year scores so that these now have standard normal distributions.

Table 2.7 gives the results of fitting two models, one (Model A) uses the normalised 8-year score as single predictor and the other (Model B) uses a logarithmic transformation together with its square. For Model A the addition of a squared term

Table 2.7 Variance components model applied to normalised JSP data.

Parameter	Model A	Model B
Fixed		
Intercept	0.01	−1.75
Age 8 score (x)	0.68 (0.026)	
$\log_e(x + 3.5)$		0.43 (0.21)
$[\log_e(x + 3.5)]^2$		0.80 (0.10)
Random		
σ^2_{u0} (between-schools)	0.079 (0.023)	0.082 (0.024)
σ^2_{e0} (between-students)	0.427 (0.023)	0.425 (0.023)
−2 loglikelihood	1505.7	1504.0

changes the likelihood by a very small amount, but for Model B we obtain a significant coefficient for the squared term so this remains in this model.

We see that the loglikelihood statistics differ, but we cannot in this case simply take the difference and use this to judge between models. Where the models are not nested in this way, we can use the Akaike Information Criterion (AIC), as a tool to search for the best fitting model. For each model the AIC is simply $l + 2p$, $(l = -2\log_e(\lambda))$, where p is the number of parameters fitted in the model and the model with the smallest AIC is chosen as the one which fits best. Note that there is no probabilistic interpretation here that would allow us to say whether one model is *significantl* better than the other; the AIC is simply an index for model comparison. We are fitting respectively four and five parameters so that in the present case the AIC for the log-transformed model is $1504.0 + 2^*5 = 1514.0$ and for the untransformed model it is $1505.7 + 2^*4 = 1513.7$. We see therefore that, despite the larger number of parameters in model B, model A has the smaller AIC, although there is a negligible difference between the values. Lindsey (1999) gives a discussion of the use of this criterion.

An alternative to the AIC is the Bayesian Information Criterion (BIC) which is computed as $l + \log_e(N^*)p$, $(l = -2\log_e(\lambda))$. The term N^* is the effective sample size. In a multilevel model it is not clear what the effective sample size should be, and the total number of higher level units is often used as an approximation. Raftery (1995) provides a discussion. For both criteria we note that the comparisons must be based on the same sample of units.

2.12.4 Inferences for residual estimates

In our JSP variance components analysis we estimated level 2 residuals, one for each school. In studies of school effectiveness, one requirement is sometimes to try to identify schools with residuals which are substantially different. From a significance testing standpoint, we will often be interested in the null hypothesis that school A has a smaller residual than school B against the alternative that the residual for school A is larger than that for school B (ignoring the vanishingly small probability that they are equal). In the case when a standard significance test accepts the alternative hypothesis (at a chosen level, say $\alpha\%$) of some difference against the null hypothesis of no difference, this is equivalent to accepting one of the alternatives (A > B, A < B) at the same level of significance and we shall use this interpretation. Alternatively we may construct a $(100 - \alpha)\%$ confidence interval for the difference between two schools means and see whether it contains zero, is less than or is greater than zero.

Where we can identify two particular schools then it is straightforward, using the results of Appendix 2.1 to construct a confidence interval for their difference or carry out a significance test. Often, however, the results are made available to a number of individuals, each of whom is interested in comparing their own schools of interest. This may occur, for example, where policy makers wish to select a few schools within a small geographical area for comparison, out of a much larger study. In the following discussion, we suppose that individuals wish to compare only pairs of schools, although the procedure can be extended to multiple comparisons of three or more residuals. Further details are given by Goldstein and Healy (1995).

Consider the JSP data where we have 48 estimated residuals together with their comparative standard errors. Since the sample size is fairly large, we shall also assume that these estimates are uncorrelated.

First, we order the residuals from smallest to largest. We construct an interval about each residual so that the criterion for judging statistical significance at the $\alpha\%$ level for any pair of residuals is whether their confidence intervals overlap. For example, if we consider a pair of residuals with a common standard error (s.e.), and assuming normality, the confidence interval width for judging a difference significant at the 5 % level is given by $\pm 1.39(\text{se})$, where $1.39 = 1.96/\sqrt{2}$.

The general procedure defines a set of confidence intervals for each residual as

$$\hat{u}_i \pm c(\text{se})_i \tag{2.17}$$

For each possible pair of intervals, (2.17), there is a significance level associated with the overlap criterion, and the value $\boldsymbol{c}$ is determined so that the average, over all possible pairs is $\alpha\%$. A search procedure can be devised to determine $\boldsymbol{c}$. When the ratios of the standard errors do not vary appreciably, say by not more than 2:1, the value 1.4 can be used for $\boldsymbol{c}$. As this ratio increases so does the value of $\boldsymbol{c}$. In the present case only 2 of these ratios are greater than 2 and we have used the common value of 1.4.

The results are presented as the 'caterpillar plot' in Figure 2.14. As is clear, apart from some of the extreme intervals, each interval overlaps with most of the other intervals. If we wished the basic comparison to take place among triplets of schools, then using the results of Section 2.12.1 we replace the normal upper 2.5 % value of 1.96 by $\sqrt{\chi^2_{2,(0.05)}} = 2.45$, since we assume that the residuals are approximately independently distributed. This will give a similar display but with intervals 25 %

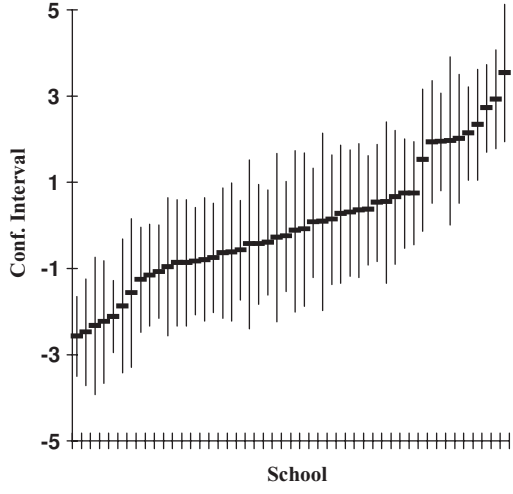


Figure 2.14 Pairwise overlap intervals for JSP school residuals.

wider and will maintain the average significance level at 5 % for all comparisons within sets of just three schools. In reality the complete set of schools typically will be compared in overlapping subsets of different sizes, and a common value for $\boldsymbol{c}$ can be determined by averaging over all such possibilities.

In other situations, we may be interested simply in judging each school against the average for all schools. In this case conventional $\alpha\%$ intervals for the residuals can be constructed and inferences based upon whether they include zero (the mean of the residuals).

We must emphasise that the above formulae should be used where a predetermined single or pairwise comparison is to be made. Suppose we wish to make a judgement, for example, about all the pairwise comparisons that are significantly different from zero. Then we must adopt a multiple comparisons procedure that takes account of the fact that under the null hypothesis all the true differences being tested are zero, so that the repeated application of a hypothesis test at significance (probability) level α will result in a probability that at least one test is significant will generally be greater than α. If the tests are independent, this probability will be $1 - (1 - \alpha)^s$, where s is the number of tests. A standard method for handling this situation to ensure an overall level of α is to use a 'Bonferroni' procedure that involves judging the significance of each test at the level α/s. This procedure, however, tends to lack power. A more powerful procedure, the sequential Bonferroni, (Holm, 1979) is first to compute the separate test p-values and rank them $p_1 < p_2 \ldots < p_s$. Then for the j-th one the null hypothesis is rejected if $p_j \leq \alpha/(s - j + 1)$, so that apart from the smallest p-value, we use a larger criterion value than the standard Bonferroni procedure and will therefore tend to produce a greater number of significant comparisons. Another popular procedure is the false discovery rate (FDR) method (see Finner and Gontscharuk, 2009). This controls the proportion of 'false discoveries', that is the expected proportion of wrongly rejected comparisons to be less than or equal to a pre-assigned value α^*. In a standard version of this procedure we find the largest value j, say k, such that $p_j \leq \frac{\alpha^* j}{s}$, and reject all hypotheses for which $j \leq k$. For more than two comparisons this will produce a greater number of significant comparisons than the sequential Bonferroni if we choose $\alpha = \alpha^*$. A useful discussion of methods is given by Afsharthus and Wolf (2007).

Presentations such as that in Figure 2.14 are useful for conveying the inherent uncertainty associated with estimates for individual level 2 (or higher) units, where the number of level 1 units per higher level unit is not large. This uncertainty in turn places inherent limitations upon such comparisons. See Goldstein and Spiegelhalter (1996) and Leckie and Goldstein (2009) for a detailed discussion of these issues in health and education.

2.13 Bayesian estimation using Markov Chain Monte Carlo (MCMC)

We now look at Bayesian models for multilevel data using MCMC methods. These incorporate prior distribution assumptions and, based upon successively sampling

from posterior distributions of the model parameters, yield a 'chain' which can then be used for constructing point and interval estimates. The two most common procedures in use are 'Gibbs sampling' and 'Metropolis-Hastings sampling'. We shall illustrate how these work for some basic models, with examples. For more details about MCMC methodology see, for example, Gilks *et al*. (1996). In later chapters we will describe the use of MCMC methods as the basis for fitting more complex models.

All MCMC algorithms are iterative and at each iteration they are designed to produce a sample from the joint posterior (multivariate) distribution of the components or parameters of the model. These parameters will be regression coefficients, covariance matrices, residuals etc. After a suitable number of iterations, we obtain a sample of values from the distribution of any parameter of set of parameters which we can then use to derive any desired distribution characteristic such as the mode, mean, covariance matrix, etc.

We outline the procedures for a general 2-level variance components model

$$y_{ij} = (X\beta)_{ij} + u_j + e_{ij}, \quad \text{var}(e_{ij}) = \sigma_e^2, \quad \text{var}(u_j) = \sigma_u^2 \tag{2.18}$$

In the Bayesian formulation of this model we combine *prior* information about the fixed and random parameters, with the likelihood based on the data. These parameters are regarded as random variables described by probability distributions, and the prior information for a parameter is incorporated into the model via a *prior distribution*. After fitting the model, the distribution produced is known as the *posterior distribution*. Formally we write for the posterior distribution $p(\theta|y)$

$$p(\theta|y) \propto L(y;\theta)p(\theta)$$

where θ represents the unknown parameters, y represents the observed responses and $L(y;\theta)$, $p\theta)$ respectively are the likelihood and the prior distribution for θ. Here we shall assume independent prior distributions for each parameter. Given a chain from the posterior distribution we can construct useful summaries, for example in the form of *point estimates* such as the mean or mode, or for quantiles such as the upper and lower $(\alpha/2)\%$ points of the distribution for a single parameter which results in an *interval estimate*. The Bayesian interpretation therefore differs from the classical frequentist interpretation where a *confidenc interval* is constructed so that the single true population mean lies within such an interval with probability $(100 - \alpha)\%$. We shall not make a rigid distinction between these interpretations when describing results which typically can be viewed from either standpoint.

The topic of the choice of priors, and in particular whether and when to use *informative* priors in a Bayesian fashion, is a vast one. A useful introduction is Draper (2002), and from time to time we will refer to the use of informative priors, for example, when discussing meta analysis (in Chapter 3) and missing values (in Chapter 16). Our default assumption, however, is that *diffuse* priors are used, in an attempt to base inference solely on the data and the results of doing this will typically be similar to the results from IGLS and RIGLS estimation. At the end of this chapter we shall see how MCMC methods can be adapted to obtain (approximate) maximum likelihood estimates.

2.13.1 Gibbs sampling

MCMC estimation is iterative and each iteration produces a set of 'current' parameter values, and after convergence the sequence of these, under general conditions, can be considered as a serially correlated random draw from the joint posterior distribution of the parameters. MCMC algorithms proceed at each iteration, by considering each component in turn and generating a random sample from the distribution of that component assuming the current values of the remaining components. The detailed steps are given in Appendix 2.5 where the choice of prior distributions is also considered. Here we outline the steps briefly. It is assumed that we have some starting values, possibly derived from a preliminary IGLS estimation.

In Gibbs sampling for the variance components model (2.18), the following steps occur at iteration t. Sampling of each set of parameters is conditional on current values of the others, the priors and the data.

Step 1
Sample a new set of fixed effects (β).
Step 2
Sample a new set of level 2 residuals $\{u_j\}$.
Step 3
Sample a new level 2 variance.
Step 4
Sample a new level 1 variance.
Step 5
Compute the level 1 residuals by subtraction. Note that the order of these steps is not important.

The first few values from the MCMC chain will usually be discarded, the 'burn in' phase, and inference will be based upon the set obtained after the chain has converged to the joint posterior distribution. The chain of parameters is treated as a (serially correlated) random sample from the full joint posterior distribution of the parameters. Typically, the mean of the chain for each parameter is chosen as equivalent to the traditional point estimate and the standard deviation as an estimate of spread, equivalent to the standard error. The modal values (constructed from a kernel density plot of the chain values), which correspond more closely to the maximum likelihood estimates when diffuse or uninformative priors are used, can also be useful. Distribution quantiles can be computed to provide interval estimates, and these can be obtained from the chain either by ranking the chain values and selecting the values corresponding to the desired quantiles, or by estimating the probability density distribution of the chain values by a suitable nonparametric smoothing method such as kernel density estimation (Silverman, 1986; Abdous and Berlinet, 1998). We can also form derived chains for functions of parameters. For example, the variance partition coefficient may be of interest and we can form a chain for this simply by calculating the appropriate ratio of the level 2 variance to the sum of the level 2 and level 1 variances at each iteration and storing these. We shall discuss these procedures in more detail in the examples below.

2.13.2 Metropolis-Hastings (MH) sampling

Gibbs sampling can be used whenever we can write down analytically the conditional posterior distributions for each parameter or group of parameters in turn. When this cannot be done easily alternative approaches can be used. Here we describe briefly Metropolis-Hastings sampling; for details of an alternative 'adaptive rejection sampling' (see Gilks and Wild, 1992).

An important case when we need to use MH sampling is with discrete response, or multilevel generalised linear, models (Chapter 4). The conditional distributions for the fixed parameters and the residuals involve exponential functions and these are not easy to sample from, so that another kind of sampling such as MH becomes necessary.

For one or more parameters at any given stage of the iterative cycle, MH proceeds by forming a *proposal distribution* for the parameters and sampling a new set of parameters from this distribution. A rule for accepting this new set or rejecting it is applied and if the proposed set is rejected we remain with the current values. This is repeated for each set of parameters in turn and, like Gibbs, chains for all the parameters are produced. Ideally we would like to have a proposal distribution that minimises the number of iterations for a given accuracy. One simple approach to creating a proposal distribution is to sample from a multivariate normal distribution. When this distribution is centred around the current parameter values it is known as a *random walk Metropolis sampler*. Further details of MH sampling are given in Appendix 2.5.

In practice we will often be able to use an efficient hybrid algorithm (see Browne and Draper, 2000) which is a mixture of MH and Gibbs sampling steps.

2.13.3 Convergence of MCMC chains

For the IGLS and RIGLS algorithms we can judge convergence by studying the relative change in successive parameter values for each parameter so that we can be assured that the estimate is accurate. When this change has reached a satisfactory small value (for example, 0.01) convergence is achieved and we can be fairly certain that, except in exceptional cases where a model is badly misspecified, further iterations will not alter these values. With MCMC chains we likewise wish to judge when we have sufficient accuracy for the mean or quantile estimate of each parameter. Here, however, because of the stochastic nature of the chain we cannot be certain that we have reached a final set of values that would not change with further iterations.

In fact there are two ways of judging convergence for chain means. We can use the standard deviation of the chain parameter values (which is the estimated standard error of the parameter) divided by the square root of the number of chain values scaled by a suitable factor to take account of the non-independence of successive chain values, to give an estimate of the (Monte Carlo) standard error of the parameter mean (see below). If we divide this by the mean itself, we can make a judgement when the value becomes suitably small, such as 0.01. For parameters which have values close to zero, however, this has little utility and a diagnostic based upon the number of significant digits accuracy is required. Such a diagnostic would provide an estimate of the number of significant digits for a given chain with a specified accuracy for the

result. To avoid specifying fractional significant digits this is better expressed in terms of the chain length needed to obtain, say, k significant digits with a specified relative accuracy. We shall look at one such diagnostic (Draper, 2002) in our example below.

One of the features of MCMC estimation is that the chain provides an estimate of the probability distribution for the parameter or for a set of parameters, and hence can be used to give interval estimates. As we have already mentioned, from a frequentist viewpoint these can be interpreted as confidence intervals and from a Bayesian perspective as credible intervals for the parameter value. The simplest method for computing these involves ranking the chain values and reading off the quantiles which cut the observed cumulative distribution at the required percentages, for example, 2.5 % and 97.5 %. Alternatively a smoothed estimated kernel density function can be used. As with the mean, we require to determine the accuracy of such quantile estimates and a commonly used one (Raftery and Lewis, 1992) estimates the chain length needed for a specified accuracy for a particular quantile.

One drawback with MCMC methods is the amount of real time that can be taken for satisfactory convergence. As we have already noted, the successive values generated in an MCMC chain are not independent since each one in turn is conditional (partly) on the previous one. If the correlation between successive values is very high then each additional value generates only a little extra information so that the chain will have to run for a comparatively long time. We can calculate the sample size based upon independent successive values that would give the same estimate for the standard error of the parameter mean. Thus, for example, if partial correlations of the second order and above are assumed to be negligible and the first order serial correlation is small, say less than 0.2, the variance of the parameter mean is $\text{var}(\bar{x}_p) \sim \frac{\sigma_p^2}{n}[1 + \frac{\rho_1(2-\rho_1)}{(1-\rho_1)}]$ so that the equivalent sample size is $\frac{n(1-\rho_1)}{1+\rho_1(1-\rho_1)}$. For larger first order correlations further terms $\frac{\sigma_p^2}{n}[\frac{\rho_1^2}{(1-\rho_1)}]$, $\frac{\sigma_p^2}{n}[\frac{\rho_1^3}{(1-\rho_1)}]$, etc. should be added to the approximate expression for $\text{var}(\bar{x}_p)$. It is therefore useful to monitor this autocorrelation, as we shall see in Section 2.13.5. Likewise, especially with MH sampling, the 'trace' of the chain of values should sweep across the full distribution of the parameter without staying too long at any one set of values; such traces are said to 'mix well' and we will be looking at these in an example in Section 2.13.5. It may often be important to experiment with setting different values for the MH proposals to achieve 'good' traces. It is also useful to run multiple chains, with different starting values, to check on stability.

The topic of MCMC chain diagnostics is an ongoing area of research and particular choices are not guaranteed to perform adequately in all circumstances.

2.13.4 Making inferences

We have explained how the chain values can be used to provide interval estimates for each parameter or a function of parameters. We can also study, for example, bivariate regions using the joint distribution derived from the chains.

We have seen how to obtain an index, the AIC, for comparing models which can be used when these were non-nested. For MCMC models we can use a similar

criterion, a generalisation of the AIC, which likewise allows comparison of models with different sets of parameters for a given set of responses and priors. This is known as the deviance information criterion (DIC) (Spiegelhalter *et al*., 2002) and is computed as follows.

At each iteration of the chain we compute the current value of –2 loglikelihood for the model being fitted using the current chain values, say D_i and write $\bar{D}$ for the mean value of this. Upon convergence we compute the deviance based upon the mean parameter values from the chain, say D. We then form $\bar{D} - D$ which is an estimate of the 'effective number of parameters' (p_D) in the model and the DIC statistic is simply $D + 2p_D$. In single level models with diffuse priors the DIC value will be very close to the AIC value and the effective number of parameters very close to the actual number of parameters in the model. For multilevel models, however, this will no longer hold. Instead, the calculation of the effective number of parameters includes the residuals which are treated as parameters to be estimated, but because they are constrained by the distributional assumptions made about them, their effective number is less than their actual number, as we shall see in Section 2.13.5. For a further discussion of the DIC and other measures for model comparison see Aitkin (2010).

2.13.5 An example

We shall illustrate MCMC estimation here with a simple example. In later chapters when we deal with specific models we will often be applying MCMC as well as (R)IGLS methods and where appropriate will describe any new types of MCMC steps.

We use the JSP data and the variance components model of Table 2.4 and fit a model using Gibbs sampling with a burn in of 500 iterations and a chain of 5000 iterations; we use diffuse priors, that is, uniform priors for the fixed coefficients and both the uniform and inverse Gamma priors for the variance parameters (see Appendix 2.4). The results, together with the IGLS (ML) estimates are given in Tables 2.8–2.10.

Several features emerge from these comparisons. We see that the P_D statistic is about 38; the fixed coefficients and variances give six standard parameters; the 48 schools are described by 32 effective parameters. The main difference between the ML and MCMC estimates is in the level 2 variances and here the Gamma priors produce results closer to the ML estimates and this will often be found (see also Browne, 1998). With just 48 schools we might expect that the choice of priors would have some effect on any parameters which depend on this number. By contrast the other parameters are based upon the total sample size so that the effect of a particular diffuse prior is very small. The other thing to notice is that the use of the standard error and a normality assumption in order to calculate an interval estimate may be misleading. Thus, for example, for the level 2 variance this would give a symmetric 95 % interval of (0.036, 0.136) as opposed to the one estimated from the chain of (0.047, 0.147) for the gamma prior. We also note that the MCMC

Table 2.8 IGLS and MCMC posterior mean estimates for JSP data with normalised scores with alternative (diffuse) priors for variances. Standard errors in brackets.

		Gibbs	
	IGLS	Uniform priors	Inverse Gamma priors
Fixed			
Intercept	0.129	0.130	0.130
Age 8 score	0.668 (0.026)	0.669 (0.027)	0.669 (0.027)
Gender (boys-girls)	−0.048 (0.050)	− 0.047 (0.050)	− 0.047 (0.050)
Social class (non-manual-manual)	0.138 (0.056)	− 0.139 (0.057)	− 0.138 (0.057)
Random			
Level 2			
σ^2_{u0}	0.080 (0.023)	0.093 (0.028)	0.086 (0.025)
Level 1			
σ^2_{e0}	0.422 (0.023)	0.426 (0.024)	0.426 (0.023)
Effective number of parameters (p_D)		38.7	38.1
DIC		1481.3	1481.1
Estimation time*	0.8 secs	16 secs	16 secs

*On a 600 MHz Pentium running Windows 2000 using MLwiN. MCMC burn in of 500 and main chain of 5000.

Table 2.9 MCMC estimates (selected parameters) for the model in Table 2.8 with alternative estimates for location and quantile estimates: uniform priors for variance parameters.

	Mean	Median	Mode	Standard deviation	2.5 %, 97.5 % quantiles
Age 8 score	0.669 (0.0004)	0.669	0.669	0.027	0.617, 0.721
Gender (boys-girls)	−0.047 (0.0007)	−0.047	−0.047	0.050	−0.146, 0.050
Random					
Level 2					
σ^2_{u0}	0.093 (0.0006)	0.089	0.085	0.028	0.051, 0.159
Level 1					
σ^2_{e0}	0.426 (0.0004)	0.426	0.415	0.024	0.382, 0.476

Table 2.10 MCMC estimates (selected parameters) for model in Table 2.8 with alternative estimates for location and quantile estimates: inverse Gamma (0.001, 0.001) priors for variance parameters.

	Mean	Median	Mode	Standard deviation	2.5 %, 97.5 % quantiles
Age 8 score	0.669 (0.0004)	0.669	0.669	0.027	0.617, 0.721
Gender (boys-girls)	−0.047 (0.0007)	−0.047	−0.047	0.050	−0.146, −0.050
Random					
Level 2					
σ^2_{u0}	0.086 (0.0006)	0.082	0.078	0.025	0.047, 0.147
Level 1					
σ^2_{eo}	0.426 (0.0004)	0.425	0.424	0.023	0.381, 0.474

modal estimates are closer to the ML estimates, especially with the gamma prior assumption.

We also see that the means are fairly accurately estimated to judge from the MCSE values. Running the chain for longer will change the values slightly and we now look at some formal diagnostics.

Figure 2.15, produced by the MLwiN software package (see Chapter 18), is for the level 2 intercept variance. It suggests that a sample size of 5000 is about adequate of the 2.5 % and 97.5 % quantiles, but that a much longer run seems to be required for the mean to be accurate to two significant figures; however, running for 25 000 iterations does not change any values appreciably. For the quantiles this sample size determines with 95 % probability (s = 0.95) that the (95 %) interval between the quantiles should vary by no more than 100*2*0.005 = 1 percentage point (r = 0.005). These diagnostics, however, are not always reliable and it is often important to continue running a chain to see how the estimates behave.

We also see, from the top left hand box, that the chain appears to have mixed well, and the box below this shows a first order correlation of just over 0.4 with little evidence of a second order partial correlation in the middle box in the right hand column. The chain gives rise to an effective sample size of about 1800.

We now look at the variance partition coefficient. The value of this, derived from the final estimates of the variances, is 0.159 for the ML estimates, 0.179 for the uniform priors and 0.168 for the Gamma priors. If we calculate this for every iteration we find that, for the Gamma priors, the mean value is 0.167 with a standard deviation of 0.042 and a 95 % interval of (0.097, 0.263). Any function of the parameters can be 'monitored' in this way, including the residuals. Thus, for each level 2 residual we will have a chain and this allows us to place interval estimates about these. This will allow us, for example, to present a 'caterpillar graph' as in

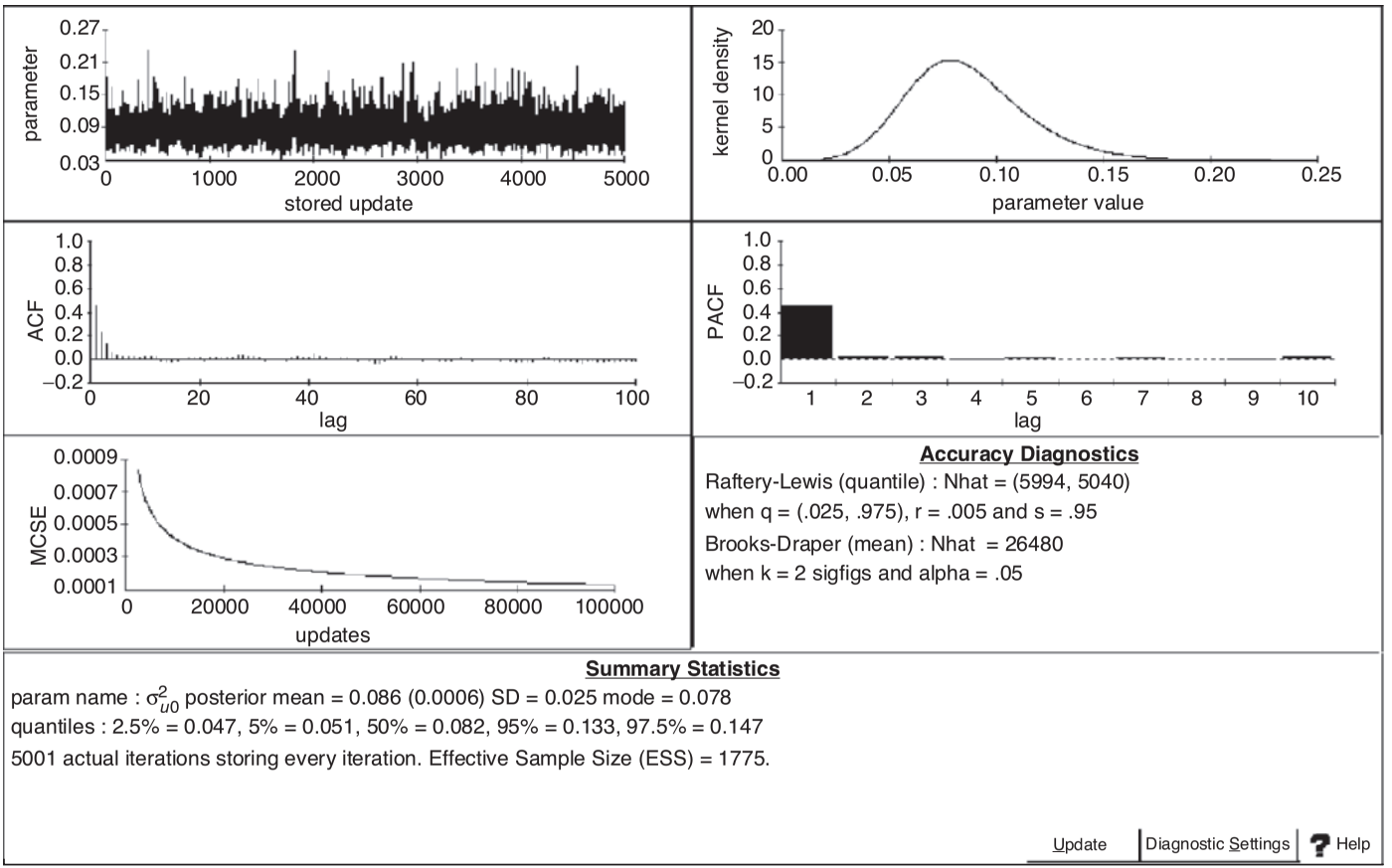


Figure 2.15 MCMC diagnostics from MLwiN.

Figure 2.12. If we wish to provide simultaneous intervals for pairwise comparisons of schools, then we can apply a scaling factor to the MCMC intervals, as discussed in section 2.12.4.

We may be more interested in presenting the *rankings* of schools rather than their means; for example, if the results are to be interpreted for the purpose of identifying which order they come in. In this case, we would form a ranking of the residuals at each iteration and evaluate the average ranking of each school, together with any interval estimate, over the chain. These final average ranks will not necessarily order the schools the same as the means, but the correlation between them will normally be high. There is a useful discussion of whether to use rankings or means in the discussion section of the paper by Goldstein and Spiegelhalter (1996). One advantage of presenting interval estimates for rankings is that it allows us to say immediately what percentage of the overall distribution for schools is covered. For example, the school with the smallest residual rank has a set of rankings for the chain of 5000 iterations extending from the lowest rank of 1 to its highest of 20 but the symmetric 95 % interval is from rank 1 to 9 only. By contrast, for the school with the highest average rank the interval is from 42 to 48. The correlation between the residual ranks and the means is 0.99.

Finally, we look at allowing for non-normality of the level 1 residuals. We shall allow them to have a general t-distribution since this is symmetrical and allows any positive or negative value. The normal distribution is the limiting case where the degrees of freedom parameter goes to infinity. Recall that we have forced the response variable to have a normal distribution by use of normal equivalent scores so we can regard the fitting of the t-distribution as a way of checking whether the normality assumption carries through to the residuals. The estimation is done by first running above Gibbs estimation with Gamma priors in MLwiN (Rasbash *et al*., 2010) and then using that program's facility for producing input to the general purpose MCMC program WINBUGS (Lunn *et al*., 2000) to fit the t-distribution (see Browne, 2008 for details). The degrees of freedom parameter for the t-distribution was also estimated, with a flat prior distribution (uniform in the range (1,10 000)). In this analysis the number of iterations was set to 5000, although often a longer chain may be needed. The results are given in Table 2.11.

The degree of freedom parameter estimate has a mean of 16.2, suggesting a fairly 'heavy-tailed' distribution, although with a wide interval estimate. Note that the estimated density for this parameter, as shown in Figure 2.16, is very skew so that the standard error should not be used to provide an interval estimate. The fixed parameter estimates do change slightly when we fit the *t*-distribution and we see that the DIC shows a modest improvement.

We shall be applying MCMC methods in subsequent chapters alongside IGLS estimation and for certain models we shall see that MCMC has particular advantages. For purposes of data driven model selection, MCMC methods can be very time consuming. Generally therefore, we can use likelihood based methods to choose one or two suitable models that fit the data and then carry out the MCMC estimation. Likewise, the model diagnostic procedures discussed in Section 2.9 will often be time-consuming and often not really feasible for use with MCMC.

Table 2.11 MCMC posterior mean estimates for the model in Table 2.8 with normal and t-distributed level 1 residuals. Standard errors in brackets (and 95 % interval for degrees of freedom). Gamma priors.

	t-distribution	normal distribution
Fixed		
Intercept	0.116	0.130
Age 8 score	0.679 (0.026)	0.669 (0.027)
Gender (boys-girls)	−0.044 (0.049)	−0.047 (0.050)
Social class (non-manual-manual	−0.134 (0.058)	−0.138 (0.053)
Random		
Level 2		
σ^2_{u0}	0.086 (0.025)	0.086 (0.026)
Level 1		
σ^2_{eo}	0.354 (0.036)	0.426 (0.023)
Effective number of parameters P_D	38.9	38.1
Degrees of freedom	16.2 (12.5) (5.9, 45.5)	
DIC	1472.6	1481.1

2.14 Data augmentation

Finally, we look briefly at a technique that can be viewed as a special kind of Gibbs sampler (Appendix 2.5). Data augmentation (Tanner and Wong, 1987) aims to provide an estimate of the full posterior distribution of the parameters in the same kind of way.

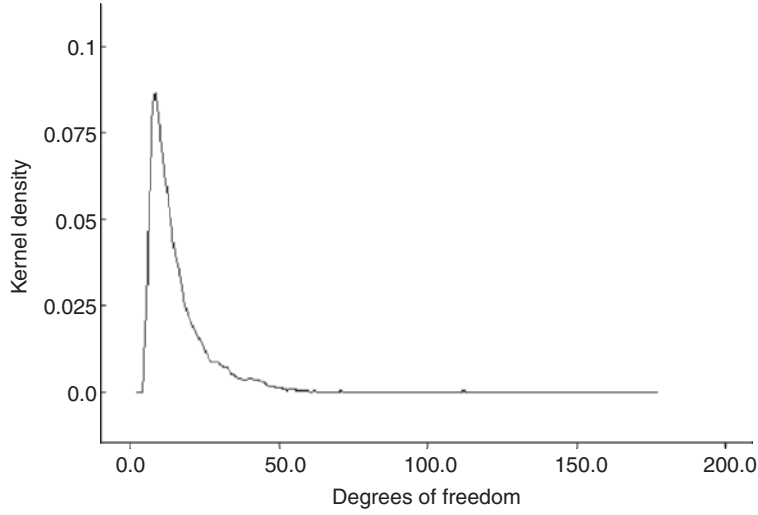


Figure 2.16 Estimated kernel density for degrees of freedom parameter in Table 2.11

The algorithm we shall consider essentially consists of two steps, the imputation (I) step and the posterior (P) step and we shall use the term IP algorithm. It is motivated, like the EM algorithm, by having 'missing data'; in multilevel models these typically are the unknown random effects but, for example, randomly missing responses can be included also. Schafer (1997) gives examples.

The IP algorithm has two steps:

Step 1

Draw a sample of the random effects given the data and current values of the model parameters. This is essentially Step 2 of the Gibbs algorithm (Appendix 2.5).

Step 2

Draw a sample from the conditional distribution of the parameters given the data and current parameter values and the random effects.

In effect, this algorithm can be regarded as a form of Gibbs sampling where all the parameters are sampled in a single step, although this step can also be broken down into smaller steps fitting random and fixed parameters in sequence. It requires a prior distribution specification, which can be diffuse, and is formally a Bayesian procedure.

An interesting application to the analysis of large cross-classified data (see Chapter 12) is given by Clayton and Rasbash (1999). They consider a set of two *parallel* IP algorithms or 'wings', one for each cross classification (A and B), with the following steps:

Step 1 for classificatio A

Given the current set of estimated residuals from classification B fit an ordinary multilevel model using, for example, IGLS.

Step 2 for classificatio A

Sample the parameters from their new conditional posterior distribution. This is essentially the combined steps 1, 3, 4 of our Gibbs sampler.

Step 3 for classificatio A

Sample a set of residuals from their new conditional posterior distribution. This is step 2 of our Gibbs sampler.

These three steps are repeated for classification B and the process run, with a suitable burn in, to produce a chain of parameters as in the standard Gibbs sampler. Clayton and Rasbash note that this method is especially suited to large datasets where the full IGLS and RIGLS methods become computationally inefficient. They also point out that in such cases, it may be unnecessary to sample the random parameters, so speeding up the computations, which is particularly important in large data problems.

Step 1 is required to set up the conditions for sampling the parameters from each wing, but a standard Gibbs sampling algorithm can be implemented by sampling residuals and parameters from all the higher level units simultaneously, although this will generally be much slower.

Appendix 2.1 The general structure and maximum likelihood estimation for a multilevel model

We illustrate the general structure using a 2-level model. We have

$$\begin{aligned} &Y = X\beta + E \\ &Y = \{y_{ij}\}, X = \{X_{ij}\}, \quad X_{ij} = \{x_{0ij}, x_{1ij}, \ldots . x_{pij}\} \\ &E = E_1 + E_2 = \{e_{ij}\}, \quad e_{ij} = e_{ij}^{(1)} + e_j^{(2)}, \\ &e_{ij}^{(1)} = \sum_{h=0}^{q_1} z_{hij}^{(1)} e_{hij}^{(1)}, \quad e_j^{(2)} = \sum_{h=0}^{q_2} z_{hij}^{(2)} e_{hj}^{(2)} \end{aligned} \tag{2.1.1}$$

We will also write simply

$$\begin{aligned} &e_{ij}^{(1)} = e_{ij}, \quad e_j^{(2)} = u_j \\ &\mathrm{Y} = \mathrm{X}\beta + \mathrm{Z}^{(2)} u + Z^{(1)} e \\ &\text{or} \\ &\mathrm{Y} = \mathrm{X}\beta + \mathrm{Z}_u u + Z_e e \end{aligned}$$

The residual matrices E_1, E_2 have expectation zero with

$$\begin{aligned} &E(E_1 E_1^T) = V_{2(1)}, \quad E(E_2 E_2^T) = V_{2(2)} \\ &E(E_1 E_2^T) = 0, \quad V_2 = V_{2(1)} + V_{2(2)} \end{aligned} \tag{2.1.2}$$

In the standard model the level 1 residuals are assumed independent across level 1 units, so that $V_{2(1)}$ is diagonal with ij-th element

$$\mathrm{var}(e_{ij}) = \sigma_{eij}^2 = z_{ij}^{(1)^T} \Omega_e z_{ij}^{(1)}$$

and Ω_e is the $(q_1 \times q_1)$ covariance matrix of the level 1 residuals.

The level 2 residuals are assumed independent across level 2 units and $V_{2(2)}$ is block-diagonal with j-th block

$$V_{2(2)j} = z_j^{(2)^T} \Omega_u z_j^{(2)}$$

and Ω_u is the $q_2 \times q_2$ covariance matrix of the level 2 residuals.

The j-th block of V_2 is therefore given by

$$V_{2j} = \oplus_i \sigma_{eij}^2 + V_{2(2)j} \tag{2.1.3}$$

where $\oplus$ is the direct sum operator.

For some of the models dealt with in later chapters, such as the time series models of Chapter 5, the requirement of independence among the residuals for the level 1 units is relaxed. In this case the first term on the right-hand side of (2.1.3) is replaced by the particular structure of $V_{2(1)}$.

For known V_2 and omitting the subscript for convenience, the generalised least squares estimate of the fixed coefficients is

$$\hat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} Y \tag{2.1.4}$$

with covariance matrix

$$(X^T V^{-1} X)^{-1}$$

For known β we form

$$Y^* = \tilde{Y}\tilde{Y}^T, \quad \tilde{Y} = Y - X\beta = E_1 + E_2 \tag{2.1.5}$$

and we have $E(Y^*) = V$. We now write

$$Y^{**} = vec(Y^*)$$

where *vec* is the vector operator stacking the columns of Y^* underneath each other. We can now write a linear model involving the random parameters, that is the elements of Ω_u, Ω_e, as follows

$$E(Y^{**}) = Z^*\theta \tag{2.1.6}$$

where Z^* is the design matrix for the random parameters. An example of such a design matrix for a simple variance components model is given in Chapter 2. We now carry out a generalised least squares analysis to estimate θ, namely

$$\hat{\theta} = (Z^{*^T} V^{*^{-1}} Z^*)^{-1} Z^{*^T} V^{*^{-1}} Y^{**}, \quad V^* = V \otimes V \tag{2.1.7}$$

where $\otimes$ is the Kronecker product. The covariance matrix of $\hat{\theta}$ is given by

$$(Z^{*^T} V^{*^{-1}} Z^*)^{-1} Z^{*^T} V^{*^{-1}} \text{cov}(Y^{**}) V^{*^{-1}} Z^* (Z^{*^T} V^{*^{-1}} Z^*)^{-1}$$

Now we have

$$Y^{**} = vec(\tilde{Y}\tilde{Y}^T) = \tilde{Y} \otimes \tilde{Y}$$

Using a standard result (see, for example, Searle *et al.*, 1992, Section 12.3) we have

$$\text{cov}(\tilde{Y} \otimes \tilde{Y}) = (V \otimes V)(I + S_N)$$

where $V \otimes V = V^*$ and S_N is the vec permutation matrix.

As Goldstein and Rasbash (1992) note, the matrix A where $Z^* = vec(A)$, is symmetric and hence

$$V^{*^{-1}} Z^* = (V^{-1} \otimes V^{-1}) vec(A) = vec(V^{-1} A V^{-1})$$

and $V^{-1} A V^{-1}$ is symmetric so that, using a standard result, we have

$$S_N V^{*^{-1}} Z^* = V^{*^{-1}} Z^*$$

and after substituting in the above expression for $\text{cov}(\hat{\theta})$ we obtain

$$\text{cov}(\hat{\theta}) = 2(Z^{*^T} V^{*^{-1}} Z^*)^{-1} \tag{2.1.8}$$

The iterative generalised least squares (IGLS) procedure (Goldstein, 1986) iterates between (2.1.4) and (2.1.7) using the current estimates of the fixed and random parameters. Typical starting values for the fixed parameters are those from an ordinary least squares analysis. At convergence, assuming multivariate normality, the estimates are maximum likelihood.

The IGLS procedure produces biased estimates in general and this can be important in small samples. Goldstein (1989a) shows how a simple modification leads to restricted iterative generalised least squares (RIGLS) or restricted maximum likelihood (REML) estimates which are unbiased. If we rewrite (2.1.5) using the *estimates* of the fixed parameters $\hat{\beta}$ we obtain

$$E(Y^*) = V_2 - X\text{cov}(\hat{\beta})X^T = V_2 - X(X^T V_2^{-1} X)^{-1} X^T \tag{2.1.9}$$

By taking account of the sampling variation of the $\hat{\beta}$, we can obtain an unbiased estimate of V_2 by adding the second term in (2.1.9), the 'hat' matrix, to Y^* at each iteration until convergence. In the case where we are estimating a variance from a simple random sample, this becomes the standard procedure for using the divisor n-1 rather than n to produce an unbiased estimate.

Full details of efficient computational procedures for carrying out all these calculations are given by Goldstein and Rasbash (1992).

The above 'plug-in' estimators for the covariance matrices ignore the fact that the parameter values θ themselves are estimates, so that they will generally be downwardly biased. This may be important in small samples and Kenward and Roger (1997) discuss corrections for the fixed effect covariance matrix estimator, using a procedure similar to that described in Appendix 2.2 for the residuals.

Appendix 2.2 Multilevel residuals estimation

2.2.1 Shrunken estimates

Denote the set of m_h distinct residuals at level h ($h>1$) in a multilevel model by

$$u^{(h)} = \{u_1^{(h)}, \ldots, u_{m_h}^{(h)}\}, \quad u_i^{(h)^T} = \{u_{i1}^{(h)}, \ldots, u_{in_h}^{(h)}\} \tag{2.2.1}$$

where n_h is the number of level h units. Since the residuals at any level are independent of those at any other level, for each residual vector we require the posterior or predicted residual estimates given by

$$\hat{u}_i^{(h)} = E(u_i^{(h)} | \tilde{Y}, V)$$

where $\tilde{Y} = Y - X\beta$. We consider the regression of the set of all residuals $u^{(h)}$ on $\tilde{Y}$ which gives the estimator

$$\hat{u}^{(h)} = R_h^T V^{-1} \tilde{Y} \tag{2.2.2}$$

where R_h is block-diagonal, each block corresponding to a level h unit and for the *j-th* block given by

$$Z_j^{(h)} \Omega_h$$

where $Z_j^{(h)}$ is the matrix of explanatory variables for the random coefficients at level h. We obtain consistent estimators by substituting sample estimates of the parameters in (2.2.2). These estimates are linear functions of the responses and their unconditional covariance matrix is given by

$$R_h^T V^{-1}(V - X(X^T V^{-1} X)^{-1} X^T) V^{-1} R_h \tag{2.2.3}$$

The second term in (2.2.3) derives from considering the sampling variation of the estimates of the fixed coefficients and can be ignored in large samples and we obtain a consistent estimator by substituting parameter estimates in

$$R_h^T V^{-1} R_h$$

Note that there are no covariances across units. Where we wish to study the distributional properties of standardised residuals for diagnostic purposes then the unconditional covariance matrix (2.2.3) should be used to standardise the estimated residuals. If, however, we wish to make inferences about the true $u^{(h)}$ for example to construct confidence intervals or test differences then we require the 'comparative' covariance matrix of $\hat{u}_j^{(h)}$ or $E[(\hat{u}^{(h)} - u^{(h)})(\hat{u}^{(h)} - u^{(h)})^T]$. The covariance matrix is

given by substituting parameter estimates in

$$S_h - R_h^T V^{-1}(V - X(X^T V^{-1} X)^{-1} X^T) V^{-1} R_h \tag{2.2.4}$$

where S_h is the block-diagonal matrix in which each block corresponds to the level h matrix Ω_h. We note that no account is taken of the sampling variability associated with the estimates of the random parameters in (2.2.3) or (2.2.4). Thus with small numbers of units, a procedure such as bootstrapping should be used to estimate these covariance matrices (Chapter 3) or a delta method approximation can be used as shown below. Skrondal and Rabe-Hesketh (2009) provide a more detailed discussion.

If we write the l-level model as

$$\begin{aligned} &y = X\beta + Zu + e \\ &Z = \{Z^{(l)}, Z^{(l-1)}, \ldots, Z^{(2)}\}, u^T = \{u^{(l)}, u^{(l-1)}, \ldots, u^{(2)}\} \\ &u \sim N(0, \Omega) \end{aligned} \tag{2.2.5}$$

then we can derive the following form for the composite residual estimator for levels 2 to l that is computationally convenient.

$$\hat{u} = (\Omega^{-1} + Z^T V_1^{-1} Z)^{-1} Z^T V_1^{-1} \tilde{y} \tag{2.2.6}$$

with comparative covariance matrix as

$$(\Omega^{-1} + Z^T V_1^{-1} Z^{-1})^{-1}$$

where $V_1 = \sigma_e^2 I_n$ and n is the number of level 1 units at the level. Corresponding to (2.2.3) we have the following additional correction factor, to be subtracted from this covariance matrix, that takes into account the sampling variation of the fixed coefficients

$$(\Omega^{-1} + Z^T V_1^{-1} Z)^{-1} Z^T V_1^{-1} X (X^T V^{-1})^{-1} X^T V_1^{-1} Z (\Omega^{-1} + Z^T V_1^{-1} Z)^{-1} \tag{2.2.7}$$

2.2.2 Delta method estimators for the covariance matrix of residuals

Consider the case of a 2-level normal model

$$y = X\beta + Z_u u + Z_e e$$

where the estimates are given by (2.2.2) and variance estimates by (2.2.4).

Using a well known equality we have

$$\text{var}(\hat{u}|y, \beta, \theta) = E[\text{var}(\hat{u}|y, \beta, \theta)] + \text{var}[E(\hat{u}|y, \beta, \theta)] \tag{2.2.8}$$

where the terms on the right hand side of (2.2.8) are regarded as functions of the model parameters and evaluated at the sample estimates. The comparative covariance matrix given by (2.2.4) is the first term in (2.2.8).

For the second term we shall use the first order approximation derived from the Taylor expansion about $E(\hat{\theta}) = \theta$, for the covariance matrix of a function, namely

$$\text{cov}[g(\hat{\theta})] \approx \left(\frac{\partial g}{\partial \theta}\right)^T \text{cov}(\hat{\theta}) \left(\frac{\partial g}{\partial \theta}\right)\Bigg|_{\hat{\theta}} \tag{2.2.9}$$

In some circumstances we may wish to have a better approximation, in which case, assuming multivariate normality, we obtain the additional contribution, evaluated at the sample estimates

$$\frac{1}{4}\left(\frac{\partial^2 g}{\partial \theta^2}\right)^T [2A_1 + A_2] \left(\frac{\partial^2 g}{\partial \theta^2}\right)\Bigg|_{\hat{\theta}}$$

$$A_1 = \{a_{ij}^2\} \quad \text{where cov}(\hat{\theta}) = \{a_{ij}\}$$
$$A_2 = aa^T \quad a = \{a_{ii}\}$$

For $\hat{u}$ as a function of the random parameters θ, we have

$$\begin{aligned} d_k^T &= \frac{\partial g}{\partial \theta_k} = \left[-\Omega_u Z^T V^{-1} \frac{\partial V}{\partial \theta_k} V^{-1} + \frac{\partial \Omega_u}{\partial \theta_k} Z^T V^{-1}\right] \tilde{y} \\ \frac{\partial \Omega_u}{\partial \theta_k} &= 0 \textit{ if } \theta_k \textit{ not at level 2.} \end{aligned} \tag{2.2.10}$$

Note that the elements of $\frac{\partial V}{\partial \theta_k}$ are just the elements of the design vector for the parameter θ_k and that

$$\frac{\partial V^{-1}}{\partial \theta_k} = -V^{-1} \frac{\partial V}{\partial \theta_k} V^{-1}$$

The row vector d_k has q elements, one for each residual at level 2 with $d = \{d_k\}$ an $t \times r_u$ matrix where t is the total number of random parameters. The second adjustment term in (2.2.8) is therefore

$$d^T \text{cov}(\hat{\theta}) d$$

This procedure for the variance of the estimated residuals is essentially equivalent to that proposed by Kass and Steffey (1989) who give an alternative derivation using the Laplace method. These authors also consider the extra adjustment term based upon the next term in the Taylor expansion as above.

Appendix 2.3 Estimation using profil and extended likelihood

Consider the following 2-level model.

$$\begin{aligned} y &= X\beta + Zu + e \\ u &\sim N(0, \Omega_u) \quad e \sim N(0, \sigma_e^2) \end{aligned} \tag{2.3.1}$$

We write twice the loglikelihood, ignoring a function of the number of observations, as

$$\begin{aligned} 2L(\theta, \beta) &= -\log|V| - tr(V^{-1}S) \\ &= -\log|V| - (Y - X\beta)^T V^{-1}(Y - X\beta) \\ S &= (Y - X\beta)(Y - X\beta)^T, \quad V = E(S) \end{aligned} \tag{2.3.2}$$

where θ is the vector of random parameters (variances and covariances) and $V = V(\theta) = E(ee^T)$. If we have a maximum likelihood (ML) estimate of β at a given value of θ then

$$\begin{aligned} 2L(\theta, \hat{\beta}) &= -\log|V| - tr(V^{-1}S) \\ &= -\log|V| - (Y - X\hat{\beta})^T V^{-1}(Y - X\hat{\beta}) \end{aligned} \tag{2.3.3}$$

is the profile likelihood (see Section 2.12.2) for the random parameters θ. The iterative generalised least squares (IGLS) algorithm alternates between maximising (2.3.3) for θ and then obtaining the conditional maximum likelihood (generalised least squares) estimate of β.

We can write the *extended likelihood*, referred to in different contexts as a penalised likelihood or an *h*-likelihood (Lee, Nelder and Pawitan, 2006), that includes the actual random effects (residuals) u treated as parameters

$$\begin{aligned} 2L(\theta, \beta, u) = &-\log|V_1| - (Y - X\beta - Zu)^T V_1^{-1}(Y - X\beta - Zu) \\ &-\log|\Omega_u| - u^T \Omega_u^{-1} u \end{aligned} \tag{2.3.4}$$

where $V_1 = \sigma_e^2 I$.

If we maximise (2.3.4) for the random effects, given (β, θ), we obtain the usual estimator (given in Appendix 2.2) which can be written conveniently as

$$\hat{u} = (Z^T \hat{V}_1^{-1} Z + \Omega_u^{-1})^{-1} Z^T \hat{V}_1^{-1}(Y - X\hat{\beta}) \tag{2.3.5}$$

Now, given estimates for θ, u, the profile likelihood for the fixed effects becomes

$$\begin{aligned} 2L(\hat{\theta}, \beta, \hat{u}) = &-\log|\hat{V}_1| - (Y - X\beta - Z\hat{u})^T \hat{V}_1^{-1}(Y - X\beta - Z\hat{u}) \\ &-\log|\hat{\Omega}_u| - \hat{u}^T \hat{\Omega}_u^{-1} \hat{u} \end{aligned} \tag{2.3.6}$$

so that a modification of the IGLS procedure is to iterate between calculating the fixed effects using (2.3.6), namely

$$\hat{\beta} = (X^T \hat{V}_1^{-1} X)^{-1} X^T \hat{V}_1^{-1} (Y - Z\hat{u})$$

which, when V_1 is diagonal, is just OLS, and then calculating the random parameters from (2.3.3) as in the standard IGLS algorithm and then the random effects from (2.3.5). Note that the OLS standard errors for the fixed coefficients based simply on (2.3.6) are incorrect. The standard errors should be computed as in Appendix 2.1.

Appendix 2.4 The EM algorithm

To illustrate the procedure, consider the 2-level variance components model

$$y_{ij} = (X\beta)_{ij} + u_j + e_{ij}, \quad \text{var}(e_{ij}) = \sigma_e^2, \quad \text{var}(u_j) = \sigma_u^2 \tag{2.4.1}$$

The vector of level 2 residuals is treated as missing data and the 'complete' data therefore consists of the observed vector Y and the u_j treated as responses. The joint distribution of these, assuming Normality, and using our standard notation is

$$\begin{bmatrix} Y \\ u \end{bmatrix} = N\left\{ \begin{bmatrix} X\beta \\ 0 \end{bmatrix}, \begin{bmatrix} V & J^T\sigma_u^2 \\ \sigma_u^2 J & \sigma_u^2 I \end{bmatrix} \right\} \tag{2.4.2}$$

This generalises readily to the case where there are several random coefficients. If we denote these by β_j we note that some of them may have zero variances. We can now derive the distribution of $\beta_j|Y$ as in Appendix 2.2, and we can also write down the normal log likelihood function for (2.4.2) with a general set of random coefficients, namely

$$\begin{aligned} \log(L) \propto -N\log(\sigma_e^2) - m\log|\Omega_u| - \sigma_e^{-2}\sum_{ij} e_{ij}^2 - \sum_j \beta_j^T \Omega_u^{-1}\beta_j \\ \Omega_u = \text{cov}(\beta_j) \end{aligned} \tag{2.4.3}$$

Maximising this for the random parameters we obtain

$$\begin{aligned} \hat{\sigma}_e^2 &= N^{-1}\sum_{ij} e_{ij}^2 \\ \hat{\Omega}_u &= m^{-1}\sum_j \beta_j\beta_j^T \end{aligned} \tag{2.4.4}$$

where m is the number of level 2 units. We do not know the values of the individual random variables. We require the expected values, conditional on the Y and the current parameters, of the terms under the summation signs, these being the sufficient statistics. We then substitute these expected values in (2.4.4) for the updated random parameters. These conditional values are based upon the 'shrunken' predicted values and their (conditional) covariance matrix, given in Appendix 2.2. With these updated values of the random parameters we can form V and hence obtain the updated estimates for the fixed parameters using generalised least squares. We note that the expected values of the sufficient statistics can be obtained using the general result for a random parameter vector θ

$$E(\theta\theta^T) = \text{cov}(\theta) + [E(\theta)][E(\theta)]^T \tag{2.4.5}$$

The prediction is known as the E (expectation) step of the algorithm and the computations in (2.4.4) the M (maximisation) step. Given starting values, based upon OLS, these computations are iterated until convergence is obtained. Convenient computational formulae for computing these quantities at each iteration can be found in Bryk and Raudenbush (2002). Meng and Dyke (1998) describe fast implementations of the EM algorithm.

Appendix 2.5 MCMC sampling

MCMC estimation proceeds, at each iteration, by considering each component in turn and generating a random sample from the distribution of that component assuming the current values of the remaining components (known as the conditional posterior distribution). Under general conditions, after the 'burn in', the resulting chain of parameter vectors can be considered as a (serially correlated) random sample from the joint posterior distribution of the parameters. To illustrate the computations consider the following variance components model.

$$y_{ij} = (X\beta)_{ij} + u_j + e_{ij}, \quad \text{var}(e_{ij}) = \sigma_e^2, \quad \text{var}(u_j) = \sigma_u^2 \tag{2.5.1}$$

2.5.1 Gibbs sampling

The Gibbs sampling algorithm proceeds through a series of steps as follows:

Step 1 *Sample a new set of fixe effects from the conditional posterior distribution*

$$p(\beta|y, \sigma_u^2, \sigma_e^2, u) \propto L(y; \beta, u, \sigma_e^2)p(\beta)$$

a suitable diffuse prior is $p(\beta) \propto 1$

$$p(\beta|y, \sigma_u^2, \sigma_e^2, u) \propto \left(\frac{1}{\sigma_e^2}\right)^{N/2} \prod_{i,j} \exp\left[-\frac{1}{2\sigma_e^2}(y_{ij} - u_j - (X\beta)_{ij})^2\right]$$

so that we sample from

$$\beta \sim MVN(\hat{\beta}, \hat{D})$$

$$\hat{\beta} = \left[\sum_{i,j} X_{ij}^T X_{ij}\right]^{-1} \left[\sum_{i,j} X_{ij}^T (y_{ij} - u_j)\right]$$

$$\hat{D} = \sigma_e^2 \left[\sum_{i,j} X_{ij}^T X_{ij}\right]^{-1}$$

Note that this is just OLS applied to the adjusted response is $\tilde{y} = (y_{ij} - (Zu)_{ij})$ and in the present case $(Zu)_{ij} = u_j$. Other assumptions for the prior are possible. Suppose we assume multivariate normality, with $p(\beta) \sim MVN(0, V)$. This is a conjugate prior, since the posterior is also multivariate normal. We sample from

$$\beta \sim MVN(\hat{\beta}^*, \hat{D}^*)$$

$$\hat{\beta}^* = \left[\sum_{i,j} X_{ij}^T X_{ij} + \sigma_e^2 V^{-1}\right]^{-1} \left[\sum_{i,j} X_{ij}^T \tilde{y}_{ij}\right]$$

$$\hat{D}^* = \sigma_e^2 \left[\sum_{i,j} X_{ij}^T X_{ij} + \sigma_e^2 V^{-1}\right]^{-1}$$

In the above and subsequent equations, the parameter terms on the right hand sides represent the current values and those on the left-hand sides the new updated ones; for simplicity we omit any indexing of the iterations where this leads to no ambiguity. The uniform prior distribution, extending over the whole real line, is 'improper' in the sense that it is not a true probability distribution. Nevertheless, for independent Normal priors the variance matrix, V, for the fixed parameters is diagonal and if the elements are very large the posterior distribution for β is equivalent to assuming uniform prior distributions. The important requirement is that the posterior distribution exists, that is it is 'proper'.

Note that we could sample each coefficient in turn but this would tend to produce more highly correlated chains compared to sampling them as a block where terms such as $[\sum_{i,j} X_{ij}^T X_{ij}]$ do not change and can be stored for use in each iteration.

Step 2 *Sample a new set of residuals*.

Each residual u_j is assumed to have a prior distribution $u_j \sim N(0, \sigma_u^2)$ which leads to the following posterior distribution, where n_j is the number of level 1 units in the j-th level 2 unit

$$p(u_j|y, \sigma_u^2, \sigma_e^2) \propto \left(\frac{1}{\sigma_e^2}\right)^{n_j/2} \prod_{i=1}^{n_j} \exp\left[-\frac{1}{2\sigma_e^2}(y_{ij} - (X\beta)_{ij} - u_j)^2\right] \times \left(\frac{1}{\sigma_u^2}\right)^{1/2} \exp\left[-\frac{1}{2\sigma_u^2} u_j^2\right]$$

so that we now sample from

$$u_j \sim N(\hat{u}_j, \hat{D})$$

$$\hat{u}_j = [n_j + \sigma_e^2 \sigma_u^{-2}]^{-1} \left[\sum_{i=1}^{n_j} (y_{ij} - (X\beta)_{ij})\right]$$

$$\hat{D} = \sigma_e^2 [n_j + \sigma_e^2 \sigma_u^{-2}]^{-1}$$

Where there are p (>1) random coefficients with explanatory variable matrix Z, this step is modified by sampling the residuals from a p-variate normal

distribution with

$$\hat{u}_j = \left[\sum_{i=1}^{n_j} Z_{ij}^T Z_{ij} + \sigma_e^2 \Omega_u^{-1}\right]^{-1} \left[\sum_{i=1}^{n_j} Z_{ij}^T (y_{ij} - (X\beta)_{ij})\right]$$
$$\hat{D} = \sigma_e^2 \left[\sum_{i=1}^{n_j} Z_{ij}^T Z_{ij} + \sigma_e^2 \Omega_u^{-1}\right]^{-1}$$

These expressions correspond to those given in Appendix 2.2.

Step 3 *Sample a new level 2 variance.*

An issue arises over the choice of a suitable diffuse prior distribution. A detailed discussion of this choice is given by Browne (1998); here we mention just two possibilities, a uniform prior, as in the case of the fixed parameters, and an inverse Gamma prior. The former choice will often tend to produce positively biased estimates. The latter is commonly used. Rather than sampling the variance directly, it is simpler to sample its inverse σ_u^{-2}, referred to as the 'precision', which is assumed to have the Gamma prior distribution $p(\sigma_u^{-2}) \sim gamma(\varepsilon, \varepsilon)$. We sample from

$$\sigma_u^{-2} \sim gamma(a_u, b_u)$$
$$a_u = (J + 2\varepsilon)/2 \quad b_u = (\varepsilon + \sum_{j=1}^{J} u_j^2/2)$$

where J is number of level 2 units.

For a uniform prior for σ_u^2 we sample from a Gamma distribution with

$$a_u = (J - 2)/2, b_u = \sum_{j=1}^{J} u_j^2/2$$

Where there are p (>1) random coefficients, we need to sample a new covariance matrix Ω_u; corresponding to the two choices for a single variance are a uniform prior over the space of covariance matrices or an inverse Wishart prior. We now sample from

$$\Omega_u^{-1} \sim Wishart(v_u, S_u)$$
$$v_u = J + v_p, S_u = \left(\sum_{j=1}^{J} u_j^T u_j + S_p\right)^{-1}$$

where u_j is the row vector of residuals for the j-th level 2 unit and the prior $p(\Omega_u^{-1}) \sim Wishart(v_p, S_p)$

A 'minimally informative' or 'maximally diffuse' choice for the prior would be to take v_p equal to the order of Ω_u and S_p equal to a value chosen to be close to

the final estimate multiplied by v_p; typically, we can use the maximum likelihood or approximate maximum likelihood estimate for S_p that is also used for the MCMC starting values. Choosing $v_p = -3$, $S_p = 0$ is equivalent to choosing a uniform prior for Ω_u.

Step 4 *Sample a new level 1 variance.*
This is similar to the procedure for a single level 2 variance. We sample from

$$\sigma_e^{-2} \sim gamma(a_e, b_e)$$
$$a_e = (N + 2\varepsilon)/2 \quad b_e = (\varepsilon + \sum_{i,j} e_{ij}^2/2)$$

For a uniform prior we can sample from a Gamma distribution with

$$a_e = (N - 2)/2, \quad b_e = \sum_{i,j} e_{ij}^2/2$$

Step 5 *Compute the level 1 residuals.*
We calculate the e_{ij} by subtraction using (2.5.1).

2.5.2 Metropolis-Hastings (MH) sampling

Consider Step 1 in 2.5.1 for the fixed coefficients. Given the current values of the fixed coefficients we form a *proposal distribution* and sample a new set of coefficients from this distribution. We then apply a rule for accepting this new set or rejecting it and remaining with the current values. Ideally, we would wish to have a proposal distribution that minimises the number of iterations for a given accuracy. We shall consider how this may be achieved after describing the basic principle.

One simple approach is to sample from a multivariate normal distribution of the following form

$$\beta_{(t)} \sim MVN(\beta_{(t-1)}, D) \tag{2.5.2}$$

where D, for example, is an estimate of the covariance matrix of the coefficient estimates, but see below for a modification. Having obtained a new set, say β^*, we need an acceptance rule and this is based upon the following ratio

$$r_t = p(\beta^*, \sigma_u^2, \sigma_e^2, u|y)/p(\beta_{(t-1)}, \sigma_u^2, \sigma_e^2, u|y) \tag{2.5.3}$$

where these probability densities have the same form as the conditional probabilities in step 1. This procedure, which conditions the proposal distribution on the current parameter values is known as a *random walk Metropolis sampler*.

We note in passing that the multivariate normal proposal distribution is symmetric in $\beta_{(t)}, \beta_{(t-1)}$ that is, $p(\beta_{(t)} = a|\beta_{(t-1)} = b) = p(\beta_{(t)} = b|\beta_{(t-1)} = a)$. Where

asymmetric distributions are used the above ratio has to be further multiplied by the so-called *Hastings ratio* $p(\beta_{(t-1)}|\beta^*)/p(\beta^*|\beta_{(t-1)})$.

We define the acceptance probability for the new set as $\alpha_t = min(1, r_t)$. If this is 1, then the new set is accepted. If it is less than 1 then it is accepted with probability α_t by, for example, generating a uniform random number in (0,1) and accepting if this is less than α_t. If the new set is not accepted then the current $(t-1)$ set is used. Thus, unlike Gibbs sampling, it is possible for the chain to stay with the same set of values for several iterations.

An efficient MH sampler attempts to avoid this as far as possible, even though the MH sampler in principle will produce satisfactory estimates if allowed to run for long enough; the key is to find a suitable proposal distribution. Gelman *et al*. (1996) suggest using *(*5.66*/d)**D, where d is the dimension of the covariance matrix D. A somewhat more satisfactory procedure (see Browne and Draper, 2000) is to specify a desired acceptance rate (for example 50 %) and adjust the proposal distribution scaling factors for each of the parameters. This can be done during a preliminary adaptive stage of the MCMC chain estimation before or during the 'burn in'.

For the variance matrix a possible proposal is to sample the inverse of the matrix from a Wishart distribution and for a single variance to sample the inverse (precision) from a Gamma distribution. We can again use an adaptive procedure to determine a scaling 'degrees of freedom' parameter. Browne *et al*. (2001) discuss various possibilities with examples, including the special case where the level 1 variance is a linear function of covariates (see Chapter 3 for a discussion of complex level 1 variation).

Generally, the variance and covariance sampling can be carried out using Gibbs and this gives an efficient hybrid algorithm (see Browne and Draper, 2000) which is a mixture of MH and Gibbs sampling steps.

2.5.3 Hierarchical centring

Consider the 3-level model

$$\begin{aligned} &y_{ijk} = (X\beta)_{ijk} + v_k + u_{jk} + e_{ijk} \\ &v_k \sim N(0, \sigma_v^2), \quad u_{jk} \sim N(0, \sigma_u^2), \quad e_{ijk} \sim N(0, \sigma_e^2) \end{aligned} \tag{2.5.4}$$

which we can write as

$$\begin{aligned} &y_{ijk} = (x'\beta')_{ijk} + u^*_{jk} + e_{ijk} \\ &u^*_{jk} \sim N(v^*_k, \sigma_u^2), \quad v^*_k \sim N(\beta_0, \sigma_v^2), \quad e_{ijk} \sim N(0, \sigma_e^2) \end{aligned} \tag{2.5.5}$$

where X', β' are the explanatory variables and coefficients excluding the intercept term.

Instead of sampling u_{jk}, v_k separately, we centre v_k on β_0 and sample the u^*_{jk} conditionally on the v_k. This procedure, with extensions to more complex models, will often greatly improve convergence (Gelfand *et al*., 1995; Browne *et al*., 2009).

2.5.4 Orthogonalisation of the explanatory variables and parameter expansion

It is common practice in fitting models with explanatory variables that are polynomials, to transform the successive polynomial powers to orthogonal variables. This use of orthogonal polynomials can be extended to any set of explanatory variables by forming an equivalent set of orthogonal variables. This is readily done using standard algorithms such as the Gram-Schmidt procedure. In the standard case where we have diffuse uniform priors for the fixed parameters the analysis then proceeds using the orthogonal variables which are finally back-transformed to the original variables together with the corresponding parameter estimates. In some applications this procedure can result in considerably improved chain mixing for the fixed parameters.

Another procedure for speeding up MCMC estimation is known as parameter expansion. This involves introducing extra parameters into the model so that it is in fact over-parameterised and not every individual parameter can then be identified. However, it is still possible to identify the functions of parameters that are required for interpretation. Thus, for example in cases where we have a variance parameter close to zero we might replace a random effect u_j by λv_j. Where λ and σ_v^2 are sampled by the chain, they cannot be separately identified, but we are able to identify the parameters of interest, namely $u_j = \lambda v_j$ and also $\sigma_u^2 = \lambda^2 \sigma_v^2$.

Browne *et al*. (2009) give examples of both these techniques with a discussion.

3

3-level models and more complex hierarchical structures

3.1 Complex variance structures

In all the models of Chapter 2 we have assumed homoscedasticity, namely, that a single variance describes the random variation at level 1. At level 2 we introduced a more complex variance structure, by allowing regression coefficients to vary across level 2 units. The modelling and interpretation of this complex variation was in terms of randomly varying coefficients. An alternative way of looking at this structure, however, is that the level 2 variance is a function of the explanatory variables that have a random coefficient at level 2. Thus, for the model of Table 2.1, if we add a random coefficient for the 8-year score to give the model

$$\begin{aligned} y_{ij} &= \beta_0 + \beta_1 x_{1ij} + (u_{0j} + u_{1j} x_{1ij} + e_{0ij}) \\ \text{var}(e_{0ij}) &= \sigma^2_{e0} \end{aligned} \tag{3.1}$$

the level 2 variance is given by

$$\text{var}(u_0 + u_1 x_{1ij}) = \sigma^2_{u0} + 2\sigma_{u01} x_{1ij} + \sigma^2_{u1} x^2_{1ij} \tag{3.2}$$

This is a quadratic function of x_1 and if we had further random coefficients the variance would be a function of all of the associated explanatory variables. This function is plotted in Figure 3.1. Although it does not occur here, the level 2 variance function can sometimes become negative within the range of the data. This is due to

Multilevel Statistical Models: 4th Edition Harvey Goldstein

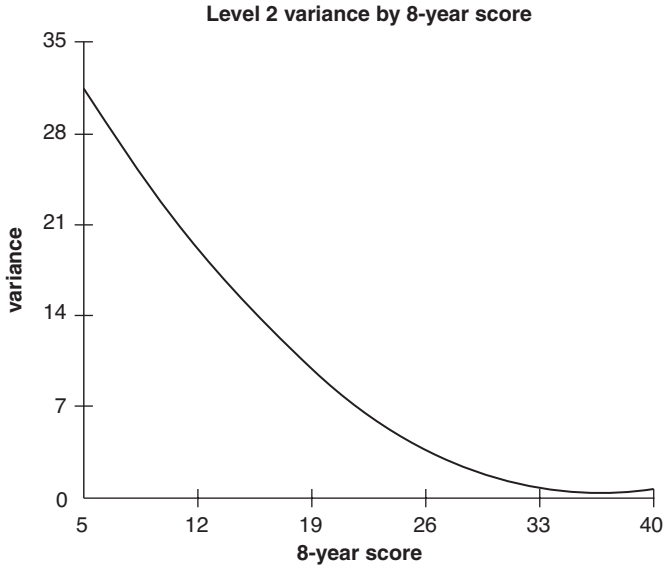


Figure 3.1 Level 2 variance function.

sampling variation since there is no restriction in the standard IGLS estimation that forces the variance function to be non-negative. One way of avoiding this problem is to have a nonlinear function such as $e^{-f(x)}$ where $f(x)$is a polynomial such as in (3.2); we deal with this case in Chapters 9 and 17.

Now we look at how we can model the variation at level 1 as a function of explanatory variables and how this can give substantively interesting interpretations.

In Figure 2.1, we saw that the level 1 residual variation appeared to decrease with increasing values of the explanatory variable, 8-year maths score.

Since we shall now consider several random variables at each level, the notation used in Chapter 2 needs to be extended. Extending the notation in (2.6) for the total level 2 random effect, we write

$$u_j = \sum_{h=0}^{l_2} u_{hj} z_{hij} \tag{3.3}$$

where there are l_2 random effects each with an explanatory variable (z_{hij}) where, by convention, z_{0ij} refers to the constant (= 1) defining a basic or intercept variance term. We can also write a similar expression for the level 1 random effect as

$$e_{ij} = \sum_{h=0}^{l_1} e_{hij} z_{hij} \tag{3.4}$$

Table 3.1 JSP mathematics data with level 1 variance a quadratic function of 8-year score measured about the sample mean. Model A with original scale; Models B and C with normal score transform of 11-year score and original 8-year score measured about its mean.

Parameter	Estimate(s.e.) Model A	Estimate (s.e.) Model B	Estimate (s.e.) Model C
Fixed:			
Intercept (x_0)	31.2	0.13	0.14
8-year score (x_1)	0.99 (0.29)	0.097 (0.004)	0.096 (0.004)
Gender (boys-girls)	−0.35 (0.26)	−0.04 (0.05)	−0.03 (0.05)
Social class (non-manual-manual)	0.74 (0.29)	0.16 (0.06)	0.16 (0.06)
School mean 8-year score	0.02 (0.11)	−0.008 (0.02)	
8-year score x school mean 8-year score	−0.02 (0.01)	0.0006 (0.02)	
Random			
Level 2			
σ^2_{u0}	2.84 (0.88)	0.084 (0.024)	0.086 (0.024)
σ_{u01}	−0.17 (0.07)	−0.0024 (0.0015)	−0.0030 (0.0015)
σ^2_{u1}	0.012 (0.007)	0.00018 (0.00016)	0.00021 (0.00016)
Level 1			
σ^2_{e0}	16.5 (1.02)	0.413 (0.029)	0.412 (0.022)
σ_{e01}	−0.90 (0.02)	−0.0032 (0.0017)	
σ^2_{e1}	0.06 (0.02)	0.0000093 (0.00041)	

In both cases, the z_{hij} may be either level 1 or level 2 defined explanatory variables. In the standard case we considered in Chapter 2 there is a single variable $z_{0ij} = 1$ and an associated level 1 variance σ^2_{e0} as in (3.1).

Suppose now that we introduce an extra random term, $e_{1ij}x_{1ij}$ into (3.1) to give

$$\begin{aligned} y_{ij} &= \beta_0 + \beta_1 x_{1ij} + (u_{0j} + u_{1j}x_{1ij} + e_{0ij} + e_{1ij}x_{1ij}) \\ \text{var}(e_{ij}) &= \text{var}(e_{0ij} + e_{1ij}x_{1ij}) = \sigma^2_{e0} + 2\sigma_{e01}x_{1ij} + \sigma^2_{e1}x^2_{1ij} \end{aligned} \tag{3.5}$$

so that we have made the level 1 variance a quadratic function of x_1. Table 3.1 shows the results of fitting this model and Figure 3.2 shows the level 1 variance function for model A.

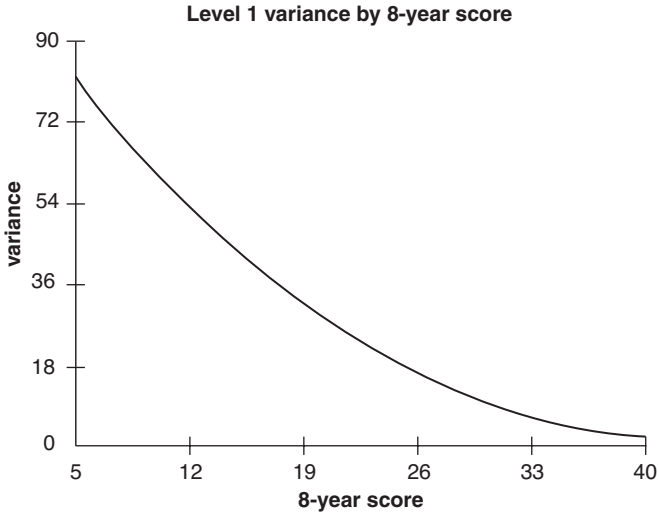


Figure 3.2 Level 1 variance function.

This is just the form of relationship we would expect given the 'ceiling' effect for the 11-year score. While this specification of the level 1 variance as a polynomial function is analogous to the specification for the level 2 variance, we do not have an alternative interpretation in terms of random slopes as we did at level 2. In other words, the parameter σ^2_{e1} cannot be interpreted as the variance of a coefficient; it is simply the coefficient of the quadratic term describing the level 1 *complex variance function.* Note that, although we have retained the same notation for the parameters of the variance function, they no longer are to be regarded as variances or covariances of measured or unmeasured variables. Of course, where a coefficient is made random at a level higher than that at which the explanatory variable itself is defined, then the resulting variance (and covariance) *can* be interpreted as the between-higher-level unit variance of the within-unit relationship described by the coefficient. When a coefficient is made random at the same or lower level than its corresponding explanatory variable then the interpretation has to be in terms of a complex variance function.

We can consider other functions, for example a linear function of x_1. Thus, if we write

$$\begin{aligned} y_{ij} &= \beta_0 + \beta_1 x_{ij} + (u_{0j} + u_{1j}x_{1ij} + e_{0ij} + e_{1ij}x_{1ij}), \\ \text{var}(e_{0ij}) = \sigma^2_{e0}, &\quad \text{var}(e_{1ij}) = 0, \quad \text{cov}(e_{0ij}e_{1ij}) = \sigma_{e01} \end{aligned} \tag{3.6}$$

the level 1 variance is now the linear function $\sigma^2_{e0} + 2\sigma_{e01}x_{1ij}$.

This device of constraining a 'variance' parameter to be zero in the presence of a nonzero 'covariance' is used to obtain the required variance structure. Thus it is only the specified *function* of the random parameters in expressions such as (3.6) which has

an interpretation in terms of the level 1 variances of the responses y_{ij}. Furthermore, we can allow a parameter such as σ^2_{e1} to be negative, so long as the total level 1 variance remains positive within the range of the data. For convenience, however, we will continue to use the notation $\text{var}(e_{1ij}) = \sigma^2_{e1}$ etc. As we have already seen, Model A shows a clear quadratic relationship which is highly statistically significant (the conventional chi squared with 2 degrees of freedom = 123). Furthermore, since we now have a better specified model, the level 2 correlation between the intercept and slope is now reduced to –0.91 from the value of –1.03 from Table 2.5, and with little change among the fixed part coefficients.

This procedure can be used generally at any level to define complex variation, where the random coefficients vary at the same or lower level at which the explanatory variables are defined. Thus for example, in the analyses of the JSP data in Chapter 2, we could model the average school 8-year-score, which is a level 2 variable, as having a random coefficient at level 2. If the resulting variance and covariance are nonzero, the interpretation will be that the between-school variance is a quadratic function of the 8-year score namely $\sigma^2_{u0} + 2\sigma_{u01}x_{1j} + \sigma^2_{u1}x^2_{1j}$ where x_{1j} is the average 8-year score.

The model (3.6) does not constrain the *overall* level 1 contribution to the variance in any way. In fact, when we try to fit this model to the JSP data we run into convergence problems since the linear function estimated at each iteration sometimes becomes negative within the range of the data. We see in Chapter 9 how the alternative nonlinear modelling of the variance overcomes this problem. If we use MCMC estimation we can overcome this problem by constraining the level 1 variance to be positive at each iteration, as follows.

Consider (3.5) where we have

$$\text{var}(e_{ij}) = \text{var}(e_{0ij} + e_{1ij}x_{1ij}) = \sigma^2_{e0} + 2\sigma_{e01}x_{1ij} + \sigma^2_{e1}x^2_{1ij} \tag{3.7}$$

At a given iteration t, suppose that we wish to update the parameter σ_{e01} given current values. Since the expression (3.7) is to remain positive for all observed values of x_{1ij} we require

$$\sigma_{e01} + \frac{\sigma^2_{e0} + \sigma^2_{e1}x^2_{1ij}}{2x_{1ij}} > 0$$

We can exclude the case $x_{1ij} = 0$, since then the variance is simply σ^2_{e0}. For the negative values of x_{1ij} we require

$$\sigma_{e01} < \min\left(\frac{\sigma^2_{e0} + \sigma^2_{e1}x^2_{1ij}}{-2x_{1ij}}\right)$$

and for positive values we require

$$\sigma_{e01} > \max\left(\frac{\sigma_{e0}^2 + \sigma_{e1}^2 x_{1ij}^2}{-2x_{1ij}}\right)$$

These constraints define a range of admissible values. Thus a standard MH proposal distribution can be used with these constraints defining truncation points for the distribution at each iteration. Each of the level 1 variance parameters is updated similarly, although for these cases only one truncation point is needed. The other parameters can be updated using Gibbs steps. Browne *et al*. (2002) describe the details for a normal proposal distribution.

One of the reasons for the high negative correlation between the intercept and slope at the school level may be associated with the fact that the 11-year score has a 'ceiling' with a third of the students having scores of 35 or more out of 40. A standard procedure for dealing with such skewed distributions is to transform the data, for example to normality, and this is most conveniently done by computing normal scores; that is by assigning normal order statistics to the ranked scores as described in Section 2.11; we adopt this procedure in future examples for this dataset, although, as we shall see in Chapter 7, there are other more general procedures for such variables.

The results from this analysis are given under Model B in Table 3.1. Note that the scale has changed since the response is now a standard normal variable with zero mean and unit standard deviation. We now find that there is no longer any appreciable complex variation at level 1; the chi-squared test for the joint hypothesis $\sigma_{e01} = \sigma_{e1}^2 = 0$ yields a value of 3.4 and the 5 % critical value based on the modified chi-squared statistic (chi-bar, see Section 2.8) is 5.1. Nor is there any effect of the compositional variable of mean school 8-year score; the chi-squared test for the two fixed coefficients associated with this give a value of 0.2 on 2 degrees of freedom. The reduced model is fitted in column C. The parameters associated with the random slope at level 2 remain significant (chi-bar = 7.7, P $\sim$ 0.02) and the level 2 correlation is further reduced to –0.71. Figure 3.3 shows the level 1 standardised residuals plotted against the predicted values from which it is clear that now the variance is much more nearly constant. This example demonstrates that interpretations may be sensitive to the scale on which variables are measured. It is typical of many measurements in the social and medical sciences that their scales are arbitrary and we can often justify nonlinear, but monotone, order preserving, transformations if they help to simplify the statistical model and the interpretation. In Section 2.12.3, we also showed how models which used different sets of *explanatory* variables, not necessarily nested ones, could be compared using the AIC criterion. Together with studying the effects of transformations of the response this can be used to arrive at a final model that provides a satisfactory, parsimonious, fit to the data.

We are not limited to making the variance a function of explanatory variables that appear in the fixed part of the model. A traditional, single level, example is 'regression through the origin' in which the fixed intercept term is zero while a level 1 variance associated with the intercept is fitted.

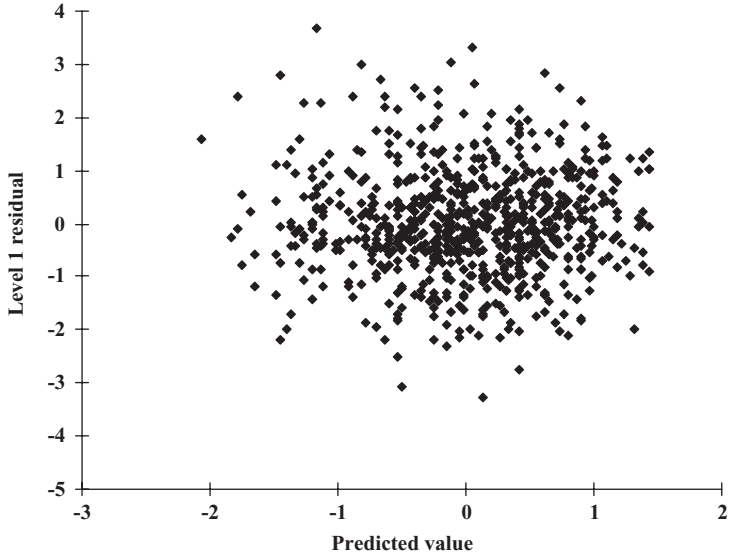


Figure 3.3 Level 1 standardised residuals by predicted values for Model C in Table 3.1.

We can consider any particular function of explanatory variables as the basis for modelling the variance. One possibility is to take the fixed part predicted value $\hat{y}_{ij}$ and define the level 1 random term as $e_{1ij}\sqrt{\hat{y}_{ij}}$, assuming the predicted value is positive, so that the level 1 variance becomes $\sigma^2_{e1}\hat{y}_{ij}$, that is, proportional to the predicted value; often known as a 'constant coefficient of variation' model that we will discuss later.

3.1.1 Partitioning the variance and intra-unit correlation

In Chapter 2, we introduced the variance partition coefficient (VPC) for a 2-level variance component model defined as

$$\tau = \frac{\sigma^2_u}{\sigma^2_u + \sigma^2_e}$$

which is the proportion of the total variance occurring at level 2. For this model this is also the correlation between two level 1 units in the same level 2 unit (the intra-unit correlation ρ). In more complex models such as (3.5), however, the intra-unit correlation is not equal to the variance partition coefficient. Thus, for (3.5) the correlation between two students (i_1, i_2) in the same school with scores x_{1i_1j}, x_{1i_2j} is

$$\frac{(\sigma^2_{u0} + \sigma_{u01}(x_{1i_1j} + x_{1i_2j}) + \sigma^2_{u1}x_{1i_1j}x_{1i_2j})}{\sqrt{\begin{array}{c}(\sigma^2_{u0} + 2\sigma_{u01}x_{1i_1j} + \sigma^2_{u1}x^2_{1i_1j} + \sigma^2_{e0} + 2\sigma_{e01}x_{1i_1j} + \sigma^2_{e1}x^2_{1i_1j}) \\ \times(\sigma^2_{u0} + 2\sigma_{u01}x_{1i_2j} + \sigma^2_{u1}x^2_{1i_2j} + \sigma^2_{e0} + 2\sigma_{e01}x_{1i_2j} + \sigma^2_{e1}x^2_{1i_2j})\end{array}}}$$

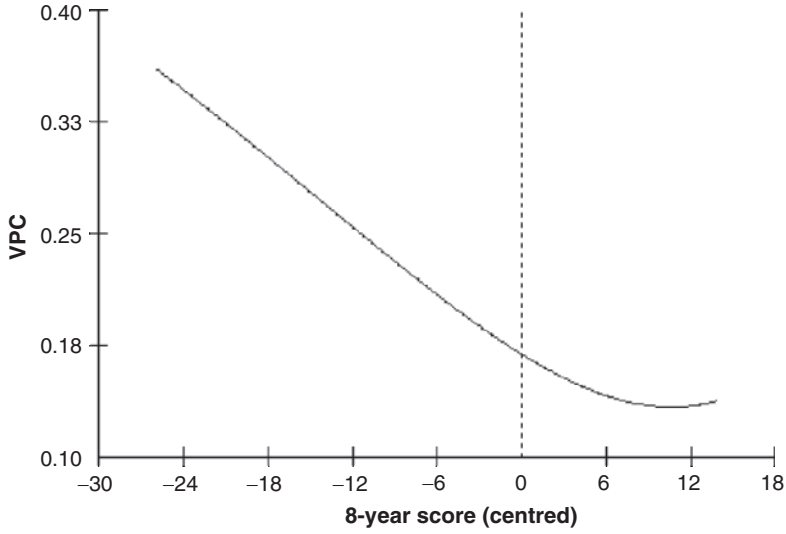


Figure 3.4 Variance partition coefficien (VPC) for Model B in Table 3.1

whereas the VPC values are

$$\frac{(\sigma^2_{u0} + 2\sigma_{u01}x_{1i_1j} + \sigma^2_{u1}x^2_{1i_1j})}{(\sigma^2_{u0} + 2\sigma_{u01}x_{1i_1j} + \sigma^2_{u1}x^2_{1i_1j} + \sigma^2_{e0} + 2\sigma_{e01}x_{1i_1j} + \sigma^2_{e1}x^2_{1i_1j})} \text{ and}$$

$$\frac{(\sigma^2_{u0} + 2\sigma_{u01}x_{1i_2j} + \sigma^2_{u1}x^2_{1i_2j})}{(\sigma^2_{u0} + 2\sigma_{u01}x_{1i_2j} + \sigma^2_{u1}x^2_{1i_2j} + \sigma^2_{e0} + 2\sigma_{e01}x_{1i_2j} + \sigma^2_{e1}x^2_{1i_2j})}$$

Figure 3.4 shows the variance partition coefficient (VPC) as a function of the 8-year score for Model B in Table 3.1. The relative variation at the school level decreases steadily with increasing 8-year score. A tentative interpretation is that the school attended is of more importance for those with below average prior achievement.

In Chapter 4, we shall consider the calculation of the VPC for models with discrete responses.

3.1.2 Variances for subgroups define at level 1

A common example of complex variation at level 1 is where variances are specific to subgroups. For example, for many measurements there may be gender or social class differences in the level 1 variation. A straightforward way to model this situation in the case of a single such grouping is by defining the following version of (3.5) for a model with different variances for children with manual and with non-manual social

Table 3.2 JSP data with normal score of 11-year maths as response. Subscript 1 refers to 8-year maths score, 2 to manual group, 3 to non-manual group, 4 to boys.

Parameter	Estimate (s.e.) Model A	Estimate (s.e.) Model B	Estimate (s.e.) Model C
Fixed			
Constant	0.13	0.13	0.13
8-year score	0.096 (0.004)	0.096 (0.004)	0.096 (0.004)
Gender (boys-girls)	−0.03 (0.05)	−0.03 (0.05)	−0.03 (0.05)
Social Class (non-manual-manual)	0.16 (0.05)	0.16 (0.05)	0.16 (0.05)
Random			
level 2			
σ^2_{u0}	0.086 (0.025)	0.086 (0.025)	0.086 (0.024)
σ_{u01}	−0.0029 (0.0015)	−0.0029 (0.0015)	−0.0028 (0.0015)
σ^2_{u1}	0.00018 (0.00015)	0.00018 (0.00015)	0.00018 (0.00015)
Level 1			
σ^2_{e0}		0.37 (0.04)	0.36 (0.04)
σ_{e02}		0.03 (0.02)	0.03 (0.02)
σ^2_{e2}	0.43 (0.03)		
σ^2_{e3}	0.37 (0.04)		
σ_{e04}			0.004 (0.02)
−2 (loglikelihood)	1491.8	1491.8	1491.7

class backgrounds.

$$
\begin{aligned}
y_{ij} &= \beta_0 + \beta_1 x_{ij} + (u_{0j} + e_{2ij} z_{2ij} + e_{3ij} z_{3ij}) \\
z_{2ij} &= 1 \text{ for manual, } 0 \text{ for non-manual} \\
z_{3ij} &= 0 \text{ for manual, } 1 \text{ for non-manual} \\
\text{var}(e_{2ij}) = \sigma^2_{e2}, &\text{var}(e_{3ij}) = \sigma^2_{e3},\ \text{cov}(e_{2ij}, e_{3ij}) = 0
\end{aligned}
$$

If we do this for Model C in Table 3.1, then we obtain the estimates in column A of Table 3.2. The estimates of the fixed parameters have changed little and the level 2 parameters are also similar. At level 1 the variance for the manual students is higher than that for the non-manual students, but not significantly so since the likelihood ratio test statistic, formed by differencing the values of (–2 log likelihood) for the model with a single level 1 variance (1493.7) and that given in Model A of Table 3.2, gives a chi-squared test statistic of 1.9 which is not significant at the 5 % level. This

Table 3.3 GCSE scores related to secondary school intake achievement.

Parameter	Estimate (s.e.) Model A	Estimate (s.e.) Model B
Fixed		
Constant	0.13	0.14
Reading score	0.50 (0.03)	0.49 (0.03)
Gender (boys-girls)	−0.19 (0.06)	−0.22 (0.06)
Social class (non-manual-manual)	−0.07 (0.06)	−0.06 (0.06)
Random		
Level 2:		
σ^2_{u0}	0.03 (0.02)	0.02 (0.01)
level 1:		
σ^2_{e0}	0.66 (0.04)	0.63 (0.04)
σ_{e01}		0.16 (0.04)
σ^2_{e1}		0.11 (0.09)
−2 (loglikelihood)	1929.5	1905.0

has the conventional 1 degree of freedom associated with it since the difference is allowed to be positive or negative, so that there is no boundary condition.

We now look at an alternative method for specifying this type of complex variation at level 1 which has certain advantages. We write

$$\begin{aligned} y_{ij} &= \beta_0 + \beta_1 x_{ij} + (u_{0j} + e_{2ij} z_{2ij}) \\ z_{2ij} &= 1 \text{ for manual, } 0 \text{ for non-manual} \\ \text{var}(e_{0ij}) = \sigma^2_{e0}, &\quad \text{var}(e_{2ij}) = 0, \quad \text{cov}(e_{0ij}, e_{2ij}) = \sigma_{e02} \end{aligned}$$

and the level 1 variance is given by $\sigma^2_{e0} + 2\sigma_{e02} z_{2ij}$ because we have constrained the variance of the manual coefficient to be zero. Thus, for manual children ($z_{2ij} = 1$) the level 1 variance is $\sigma^2_{e0} + 2\sigma_{eo2}$ and for non-manual children the level 1 variance is σ^2_{e0}. The second column in Table 3.2 gives the results from this formulation and we see that, as expected, the covariance estimate is equal to half the difference between the separate variance estimates in the first column.

Suppose now that we wish to model the level 1 variance as a function both of social class group and gender. One possibility is to fit a separate variance for each of the four possible resulting groups, using either of the above procedures. Another possibility is to consider a more parsimonious 'additive' model for the variances, as follows.

$$\begin{aligned} e_{ij} &= e_{0ij} + e_{2ij} z_{2ij} + e_{4ij} z_{4ij} \\ z_{4ij} &= 1 \text{ if a boy, } 0 \text{ if a girl} \\ \text{var}(e_{0ij}) = \sigma^2_{e0}, &\quad \text{cov}(e_{0ij} e_{2ij}) = \sigma_{e02}, \text{cov}(e_{0ij} e_{4ij}) = \sigma_{e04} \end{aligned} \tag{3.8}$$

with the remaining two variances and covariance equal to zero. Thus, (3.8) implies that the level 1 variance for a manual boy is $\sigma_{e0}^2 + 2\sigma_{e02} + 2\sigma_{e04}$ etc. The third column of Table 3.2 gives the estimates for this model and we see that there is a negligible difference in the level 1 variance for boys and girls.

We can extend such structuring to the case of multicategory variables and we can also include continuous variables as in Table 3.1. Suppose we had a 3-category variable: we define two dummy variables, say z_{5ij}, z_{6ij} corresponding to the second and third categories, just as if we were fitting the factor in the fixed part of the model. With z_{1ij} representing the continuous variable, an additive model for the level 1 random variation can be written as

$$\begin{aligned} e_{ij} &= e_{0ij} + e_{1ij}z_{1ij} + e_{5ij}z_{5ij} + e_{6ij}z_{6ij} \\ \operatorname{var}(e_{0ij}) &= \sigma_{e0}^2, \quad \operatorname{var}(e_{1ij}) = \sigma_{e1}^2, \quad \operatorname{cov}(e_{0ij}e_{1ij}) = \sigma_{e01} \\ \operatorname{cov}(e_{0ij}e_{5ij}) &= \sigma_{e05}, \quad \operatorname{cov}(e_{0ij}e_{6ij}) = \sigma_{e06} \end{aligned}$$

with the remaining terms equal to zero.

This model can be elaborated by including one or both the covariances between the dummy variable coefficients and the continuous variable coefficient, namely $\sigma_{e15}, \sigma_{e16}$. These covariances are analogous to interaction terms in the fixed part of the model and we see that, starting with an additive model, we can build up models of increasing complexity. The only restriction is that we cannot fit covariances between the dummy variable categories for a single explanatory variable. Thus, if social class had three categories, we could fit two covariances corresponding to, say, categories 2 and 3 but not a covariance *between* these categories.

Residuals can be estimated in a straightforward manner for these complex variation models. For example, from (3.8) the estimated residual for a manual boy is $\hat{e}_{0ij} + \hat{e}_{2ij} + \hat{e}_{4ij}$ where the estimates of the individual residuals are computed using the formulae in Appendix 2.2 with the appropriate zero variances.

3.1.3 Variance as a function of predicted value

The level 1 variance can be modelled as a function of any combination of explanatory variables and in particular we can incorporate the fixed coefficients themselves in such functions. A useful special case is where the function is the fixed part predicted value, $\hat{y}_{ij}$. Thus, (3.7) becomes

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + (u_{0j} + e_{0ij} + e_{1ij}\hat{y}_{ij})$$

with level 1 variance given by $\sigma_{e0}^2 + 2\sigma_{e01}\hat{y}_{ij} + \sigma_{e1}^2\hat{y}_{ij}^2$. A special case of this model is the so-called 'constant coefficient of variation model' where the two variance terms are constrained to zero, and $\hat{y}_{ij}$ remains positive. The estimation of the random parameters is straightforward: at each iteration of the algorithm a new set of predicted values are calculated and used as the level 1 explanatory variable.

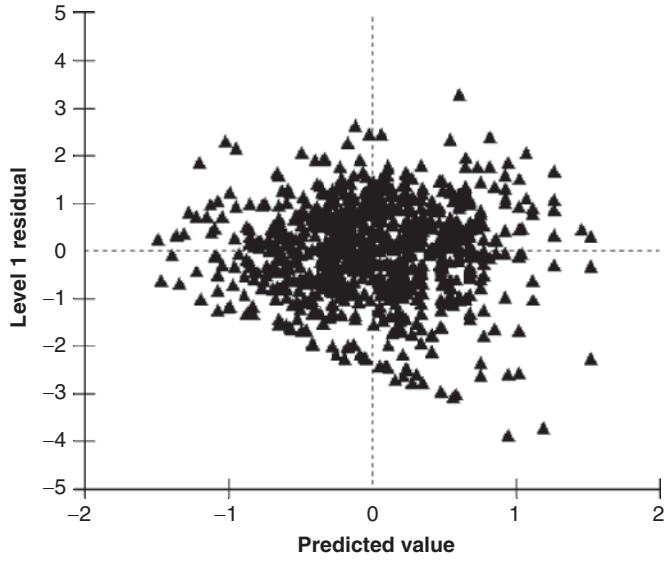


Figure 3.5 Standardised residuals for variance components analysis in Table 3.3.

Table 3.3 illustrates the use of this model where the level 1 variance shows a strong dependence on the predicted value. The data are the General Certificate of Secondary Examination (GCSE) scores at the age of 16 years of the Junior School Project students. This score is derived by assigning values to the grades achieved in each subject examination and summing these to produce a total score (see Nuttall *et al*., 1989 for a detailed description). There are 785 students in this analysis in 116 secondary schools to which they transferred at the age of 11 years. The students have a measure of reading achievement, the London Reading Test (LRT) taken at the end of their junior school and this is used as a pretest baseline measure against which relative progress is judged. Both the reading test score and the examination score have been transformed to normal equivalent deviates.

Model A is a variance components analysis and Figure 3.5 shows a plot of the standardised level 1 residuals against the predicted values. It is clear that the variation is much smaller for low predicted values.

One possible extension of the model to deal with this is the use the LRT score as an explanatory variable at level 1, so that the level 1 variance becomes a quadratic function of LRT score. This does not, however, entirely eliminate the relationship and instead we model the predicted value as a level 1 explanatory variable, and the results are presented as Model B of Table 3.3. If we now plot the standardised residuals associated with the intercept against the predicted values we obtain the pattern in Figure 3.6 from which it is clear that much of the relationship between the variance and the predicted value has been accounted for. We could go on to fit more complex functions of the predicted value, for example involving nonlinear or higher order polynomial terms.

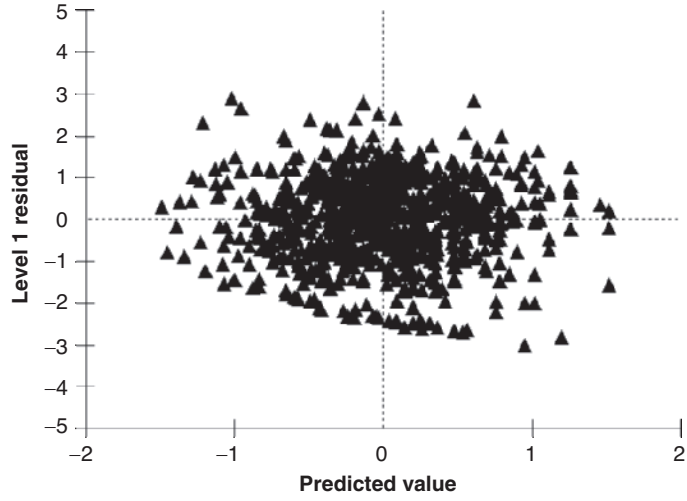


Figure 3.6 Standardised residuals with level 1 variance a function of predicted value in Table 3.3.

3.1.4 Variances for subgroups define at higher levels

The random slopes model in Table 3.1 has already introduced complex variation at level 2 when the coefficient of a level 1 explanatory variable is allowed to vary across level 2 units. Just as with level 1 complex variation, we can also allow coefficients of variables defined at level 2 to vary at level 2. Exactly the same considerations apply for categorical level 2 variables as we had for such variables at level 1 and complex additive or interactive structures can be defined.

The coefficient of a level 2 variable can vary randomly at either level 1 or level 2 or both. For example, suppose we have three types of school; all boys schools, all girls schools and mixed schools. We can allow different variances, at level 2, between boys' schools, between girls' schools and between mixed schools. We can also allow different between-student variances for each type of school.

To further illustrate complex level 1 variation and also to introduce a 3-level model, we look at another dataset, this time from a survey of social attitudes.

3.2 A 3-level complex variation model example

The longitudinal or panel data come from the British Social Attitudes Survey and cover the years 1983–86 with a random sample of 264 adults, measured a year apart on four occasions and living at the same address. This panel was a subsample of a larger series of cross-sectional surveys. The final sample was intended to be self-weighting, with each household as represented by a single person having the

same inclusion probability. A full technical account of the sampling procedures is given by McGrath and Waterton (1986). The sampling procedure was at the first stage to sample parliamentary constituencies (the primary sampling units – PSUs) with probability proportional to size of electorate, then to sample a single 'polling district' within each constituency in a similar way and finally to sample an equal number of addresses within each polling district.

Because only one polling district was sampled from each constituency, we cannot separate the between-district from the between-constituency variation; the two are 'confounded'. Likewise we cannot separate the between-individual from the between-household variation. The basic variation is therefore at two levels, between-districts (constituencies) and between-individuals (households). The longitudinal structure of the data, with four occasions, introduces a further level below these two, namely a between-occasion-within-individual level, so that occasion is level 1, individual is level 2 and district is level 3. In Chapter 5, we study longitudinal data structures in more depth, both at level 1 and higher levels.

The response variable we shall use is a scale, in the range 0–7, concerned with attitudes to abortion. It is derived by summing the (0,1) responses to seven questions and can be interpreted as indicating whether the respondent supported or opposed a woman's right to abortion with high scores indicating strong support. Explanatory variables are political party allegiance (four categories), self-assessed social class (three categories), gender, age (continuous), and religion (four categories) and year (four categories). A number of preliminary analyses have been carried out and the effects of party allegiance, social class, gender, and age, were found to be small and not statistically significant. We therefore examine the basic 3-level model, which can be written as follows

$$\begin{aligned} y_{ijk} = \beta_0 &+ (\beta_1 x_{1ijk} + \beta_2 x_{2ijk} + \beta_3 x_{3ijk}) \\ &+ (\beta_4 x_{4ijk} + \beta_5 x_{5ijk} + \beta_6 x_{6ijk}) + (v_k + u_{jk} + e_{ijk}) \end{aligned} \tag{3.9}$$

with the explanatory variables with subscripts 1–3 being dummy variables for religious categories 2–4 and those with subscripts 4–6 being dummy variables for years 1984–86. We have three variances, one at each level in the random part of the model. The response variable in the following analyses has only eight categories, with 32 % of the sample having the highest value of 7.

The response has been transformed by assigning normal scores to the overall distribution and we shall treat the response as if it was continuously distributed. Other models which retain the categorisation of the response variable are considered in Chapter 4.

Table 3.4 gives the results of fitting (3.9). The between-occasion and between-individual variances are similar. The level 3 variance is small, and the likelihood ratio chi-squared is 2.05 (compared with a value of 1.64 obtained from comparing the estimate with its standard error), neither of which is significant at the 5 % level using the chi-bar test.

Table 3.4 Repeated measurements of attitudes to abortion. Response is normal score transformation. Religion estimates are contrasted with none. Age is measured about the mean of 37 years.

Parameter	Estimate (S.E.) Model A	Estimate (S.E.) Model B	Estimate (S.E.) Model C
Fixed			
Constant	0.32	0.33	0.33
Religion:			
Roman Catholic	−0.80(0.18)	−0.80(0. 81)	−0.69(0.18)
Protestant	−0.27(0.10)	−0.26(0.10)	−0.25(0.10)
Other	−0.63(0.13)	−0.63(0.13)	−0.54(0.14)
Year:			
1984	−0.29(0.05)	−0.29(0.48)	−0.29(0.05)
1985	−0.06(0.05)	−0.07(0.05)	−0.07(0.05)
1986	0.06(0.05)	0.05(0.04)	0.05(0.04)
Age:			0.013(0.005)
Age x Roman Catholic			−0.036(0.010)
Age x Protestant			−0.014(0.007)
Age x Other			−0.023(0.008)
Random			
Level 3			
σ_v^2	0.03(0.03)	0.03(0.02)	0.03(0.02)
Level 2			
σ_u^2	0.37(0.04)		0.34(0.04)
Level 1			
σ_{e0}^2	0.31(0.02)	0.21(0.08)	0.21(0.03)
σ_{e01}		0.11(0.05)	0.10(0.04)
σ_{e02}		0.03(0.16)	0.03(0.02)
σ_{e03}		0.04(0.02)	0.04(0.02)
σ_{e04}		0.05(0.02)	0.05(0.02)
σ_{e05}		0.05(0.02)	0.05(0.02)
σ_{e06}		0.00(0.02)	0.00(0.02)
−2 (loglikelihood)	2233.5	2214.2	2198.7

For the religious differences we have $\chi_3^2 = 33.7$ for the overall test with all those having religious beliefs being less inclined to support abortion, the Roman Catholic and other religions being least likely of all. The Roman Catholic and other religions are significantly less likely than the Protestants to support abortion. The simultaneous test (3 d.f.) chi-squared statistics respectively are 9.7 and 9.0 (P = 0.03).

For the year differences we have $\chi_3^2 = 59.7$ and simultaneous comparisons show that in 1984 there was a substantially less-approving attitude towards abortion. It

is likely that this is an artefact of the way questions were put to respondents.[1] No significant interaction exists between religion and year.

We now look at elaborating the random structure of the model. At level 1 we fit an additive model, as in Section 3.1, for the categories of religion and for year. Year is the variable defining level 1, but religion is a time invariant variable defined at level 2 and is an example of a higher level variable used to define complex variation at a lower level.

The results are given as Model B in Table 3.4. For year we obtain $\chi^2_3 = 8.3$ (P = 0.04) and for religion $\chi^2_3 = 11.0$ (P = 0.01). We note again that conventional chi-squared tests are used since the differences being tested can be positive or negative. There is a greater heterogeneity within the Roman Catholics, from year to year, and within the other religions than within Protestants and those with no religion. The addition of these variances to the model does not change substantially the values for the other parameters.

Fitting complex variation at level 2 (between individuals) and level 3 (between districts) does not yield statistically significant effects, although there is some suggestion that there may be more variation among Roman Catholics.

For the final analysis, we look again at the fixed part and explore interactions. None of the interactions have important effects except for that of age with religion, although age on its own had a negligible effect. We see from Model C that those with no religion show an increasing approval of abortion with age, whereas the Roman Catholics and to a smaller extent other religions show a decreasing approval with age. The overall chi-squared for testing the interactions is 16.1 with three degrees of freedom (P = 0.001). In Chapter 4, we give an example of modelling level 2 variation to study segregation patterns across schools.

3.3 Parameter constraints

In the example of the previous section, some of the fixed and random parameters for year and religious groups were similar. This suggests that we could fit a simpler model by forcing or 'constraining' such parameters to take the same values resulting in a decrease in the standard errors in the model. We illustrate the procedure using the fixed part estimates for the abortion attitudes data.

We consider the general constraint for the fixed parameters in the form of n linear constraint functions defined by $C^T\beta = k$, where C is a $(p\ x\ n)$ constraint matrix and k is a vector, and these can have quite general values for their elements.

Suppose that in Model C of Table 3.4 we wished to constrain the main effects and interaction terms of the Roman Catholic and Other religions to be equal. This

[1] In 1984, seven questions making up the attitude scale were put to respondents in the reverse order, that is with the most 'acceptable' reasons for having an abortion (for example, as a results of rape) coming first. This illustrates an important issue in surveys of all kinds which collect data for comparisons over time, namely to maintain the same questioning procedure.

implies $n = 2$ constraint functions, and we have

$$C^T = \begin{pmatrix} 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 \end{pmatrix}$$

$$k = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

which implies $\hat{\beta}_1 = \hat{\beta}_3,\ \hat{\beta}_8 = \hat{\beta}_{10}$.

The constrained estimator of β is

$$\begin{aligned} \hat{\beta}^c &= \hat{\beta} - LC(C^T LC)^{-1}(C^T\hat{\beta} - k) \\ L &= (X^T\hat{V}^{-1}X)^{-1} \end{aligned} \tag{3.10}$$

where $\hat{\beta}$ is the unconstrained estimator. The covariance matrix of the constrained estimator is MLM^T where

$$M = I - LC(C^T LC)^{-1}C^T.$$

There is an analogous formula for constrained random parameters.

Using the above constraints for Model C in Table 3.4, the random parameters are little changed, the main effects for Roman Catholic and Other religion become –0.57 and the interaction terms become –0.026 and the remaining main effects are virtually unaltered. The standard errors, as expected, are smaller: 0.121 for the main effect estimate and 0.007 for the interaction.

In addition to linear constraints, we can also apply nonlinear constraints. To illustrate the procedure we consider the analysis in Table 2.5, where the estimated correlation between the slope and intercept was –1.03. To constrain this to be exactly –1.0, after each iteration of the algorithm we compute the covariance as a function of the variances to give this correlation. Thus, after iteration t we compute $\sigma_{u01}^{t+1} = \hat{\sigma}_{u0}^t\hat{\sigma}_{u1}^t$ and then constrain the covariance to be equal to this value, a linear constraint, for iteration $t+1$. This procedure is repeated until convergence is obtained for the unconstrained values. For more general nonlinear constraints we may require several such constraints to apply simultaneously.

If we constrain the model of Table 2.5 to give a correlation of –1.0 we find that the fixed effects and the level 1 variance are altered only slightly, with a small reduction in standard errors. The level 2 parameters, however, are reduced by about 50 % and are closer to those in Model A of Table 3.1 where the estimated correlation is –0.91.

We can also temporarily constrain values during the iterative estimation procedure if convergence is difficult or slow. Some parameters, or functions of them, can be held at current values, other parameter values allowed to converge and the constrained parameters subsequently unconstrained.

Table 3.5 Robust (sandwich) standard errors for Model A in Table 3.4.

Parameter	Estimate	Model-based s.e.	Robust s.e.
Fixed			
Constant	0.32		
Religion:			
Roman Catholic	−0.80	0.176	0.225
Protestant	−0.27	0.098	0.102
Other	−0.63	0.127	0.121
Year:			
1984	−0.29	0.048	0.050
1985	−0.06	0.048	0.061
1986	0.06	0.048	0.047
Random			
Level 3			
σ_v^2	0.03	0.030	0.020
Level 2			
σ_u^2	0.37	0.043	0.039
Level 1			
σ_{e0}^2	0.31	0.016	0.022

3.4 Weighting units

It is common in sample surveys to select level 1 units, for example, household members, so that each unit in the population has the same probability of selection. Such self-weighting samples can then be modelled using any of the multilevel models of this book. Likewise, if the model correctly specifies the population structure, non-self weighting samples can be modelled similarly: the differential selection probabilities contain no extra information for the model parameters.

In some cases, however, the sample inclusion probabilities are designed to be equal but there is differential nonresponse. It is common in such cases to assign weights to the level 1 units to compensate for the nonresponse. Another situation where weighting should be used is when the modelling of the population structure is inadequate, that is the model is misspecified. If we then wish to make inferences about population values which are robust against poor model specification we will need to weight the units. In some data analyses, we may come across values which are possible errors. Rather than excluding the units containing these, we may wish to keep them but assign them a lower weighting in the analysis in line with the extent to which we believe they are in error. In Chapter 16, we investigate in further detail the application of weighting for sample surveys where weights are used in conjunction with imputation procedures, especially where there is attrition in longitudinal surveys.

We now go on to describe the application of weights under the assumption that the weights themselves are unrelated to the random effects. Pfefferman *et al*. (1998) describe a weighted or pseudo maximum likelihood procedure where weights may be informative; in Section 3.4.2, we extend this to MCMC estimation. Pfefferman *et al*. (1998) show that the simple procedure in the Section 3.4.1 will produce acceptable results in many cases but can give biased results in some circumstances and should be used with caution.

3.4.1 Maximum likelihood estimation with weights

Consider the case of a 2-level model. Denote by w_j the weight attached to the j-th level 2 unit and by $w_{i|j}$ the weight attached to the i-th level 1 unit within the j-th level 2 unit. This allows for the possibility of having differential weights at both levels. We scale the weights so that

$$\sum_i w_{i|j} = n_j, \quad \sum_j w_j = m \tag{3.11}$$

where m is the total number of level 2 units and $N = \sum_j n_j$ the total number of level 1 units. Thus, the lower level weights within each immediate higher level unit are scaled to have a mean of unity, and likewise for higher levels. For each level 1 unit we now form the final, or composite, weight

$$w_{ij} = N w_{i|j} w_j \Big/ \sum_{i,j} w_{i|j} w_j = N w_{i|j} w_j \Big/ \sum_j n_j w_j \tag{3.12}$$

Denote by Z_u, Z_e respectively the sets of explanatory variables defining the level 2 and level 1 random coefficients and form

$$\begin{aligned} Z_u^* &= W_j Z_u, \quad W_j = diag\{w_j^{-0.5}\} \\ Z_e^* &= W_{ij} Z_e, \quad W_{ij} = diag\{w_{ij}^{-0.5}\} \end{aligned} \tag{3.13}$$

We now carry out a standard estimation but using Z_u^*, Z_e^* as the random coefficient explanatory variables rather than Z_u, Z_e.

For a 3-level model, with an obvious extension to notation, we have the following

$$\begin{aligned} &\sum_i w_{i|jk} = n_{jk}, \quad \sum_j w_{j|k} = m_k, \quad \sum_k w_k = K, \quad N = \sum_{jk} n_{jk}, \quad m = \sum_k m_k \\ &w_{ijk} = N w_{i|jk} w_{j|k} w_k \Big/ \sum_{ijk} w_{i|jk} w_{j|k} w_k, \quad w_{jk} = m w_{j|k} w_k \Big/ \sum_{jk} w_{j|k} w_k \end{aligned} \tag{3.14}$$

where K is the number of level 3 units.

We have analogous results for the random parameter estimates. Standard errors can be calculated using sandwich estimators, as described in Section 3.5.

In survey work, analysts often have access only to the final level 1 weights w_{ij}. In this case, and assuming that these final level weights have been computed as above, then for a 2-level model we can obtain the w_j by computing $w'_j = mW_j / \sum_j W_j$, $W_j = (\sum_i w_{ij})/n_j$. For a 3-level model the procedure is carried out for each level 3 unit and the resulting w'_{jk} are transformed analogously.

A similar procedure applies for multilevel generalised linear models (see Chapter 4). Here the weighted explanatory variables at levels 2 and higher are as above. For the quasilikelihood estimators (PQL and MQL) at level 1 the vector Z_e is that which defines the binomial variation. Thus, for binomial data, at level 1 a method of incorporating the weight vector is to use Z_e but to work with $w_{ij}n_{ij}$ instead of n_{ij} as the denominator.

A number of features are worth noting. For a single level model this procedure gives the usual weighted regression estimator. Secondly, suppose we set a particular level 1 weight to zero. This is not equivalent to removing that unit from the analysis in a 2-level model since the level 2 (weighted) contribution remains. Nevertheless, this weighting may be appropriate if we wish to remove the effect of the unit only at level 1, say if it were an extreme level 1 outlier. If, however, we set a level 2 weight to zero then this *is* equivalent to removing the complete level 2 unit. If we wished to obtain estimates equivalent to removing the level 1 unit we would need to set all the level 2 (random coefficient) explanatory variables *for that level 1 unit* to zero also. This is easily done by defining an indicator variable for the unit (or units) with a zero corresponding to the unit in question and multiplying all the random explanatory variables by it. We also note that when the level 2 weights are equal this procedure is equivalent to that proposed by Pfeffermann *et al*. (1998).

3.4.2 Weighted MCMC estimation

Within the standard MCMC algorithm (Appendix 2.5) we incorporate weights by forming the weighted likelihood, analogously to Pfeffermann *et al*. (1998). For the fixed effects with uniform diffuse priors this gives the posterior distribution

$$\left[\sum_{ij} w_{ij}(X_{ij}^T X_{ij})\right]^{-1} \sum_{ij} w_{ij} X_{ij}^T \tilde{y}_{ij}, \quad \tilde{y}_{ij} = y_{ij} - z_{ij}u_j$$

with covariance matrix

$$\sigma_e^2 \left[\sum_{ij} w_{ij}(X_{ij}^T X_{ij})\right]^{-1}$$

For the random effects at level 2 the posterior is

$$\left[\sum_i w_{i|j} z_{ij}^T z_{ij} + \Omega_2^{-1}\sigma_e^2\right]^{-1}\left[\sum_i w_{i|j} z_{ij}^T (y_{ij} - X_{ij}\beta)\right]$$

with covariance matrix

$$\left[\sum_i w_{i|j} z_{ij}^T z_{ij} + \sigma_e^2\Omega_2^{-1}\right]^{-1}$$

For the level 2 covariance matrix the posterior is

$$\Omega_2^{-2} \sim \mathit{Wishart}(v, S), \quad v = M - 3, \quad S = \sum_j w_j u_j^T u_j$$

and for the level 1 variance we have the posterior

$$\sigma_e^{-2} \sim \mathit{gamma}(a_e, b_e)$$
$$a_e = (N + 2\varepsilon)/2, \quad b_e = (\varepsilon + \sum_{i,j} w_{ij} e_{ij}^2/2)$$

We note that this is not equivalent to using, as before, transformed variables, Z_u^*, Z_e^* unless the level 2 weights are equal, and this is the counterpart of the result quoted above from Pfeffermann *et al*. (1998). Further details with an example are given by Goldstein *et al*. (2010).

3.5 Robust (sandwich) estimators and jacknifin

Until now we have assumed that the response variable has a normal distribution, and where the departure from normality is substantial we have considered transformations, for example using normal scores. As we saw in the abortion dataset, however, such transformations may not provide good approximations to normality, especially where the original score distribution is highly discrete or very skew. The estimates of the fixed and random parameter estimates will still be consistent when the normality assumption is untrue, but the standard error estimates cannot be used to obtain confidence intervals or to test significance except in large samples. In other cases we may have misspecified the random part of the model.

One way of attempting to deal with this problem is to develop estimators based upon alternative distributional assumptions; in Chapter 7, we adopt this approach when dealing with discrete response data. We have seen in Chapter 2 the use of a t-distribution as an alternative to the normal for the level 1 residuals and Seltzer (1993) gives another example using a t-distribution.

An alternative procedure is to modify the standard error and confidence interval estimates so that they are less dependent on distributional assumptions, of whatever

kind. One of the consequences of this is that the resulting significance tests and confidence intervals will tend be wider, or more 'conservative', than those derived under a particular distributional assumption.

Consider first the fixed part of the model and the usual IGLS estimate of the fixed parameters based upon the random parameter estimates

$$\hat{\beta} = (X^T \hat{V}^{-1} X)^{-1} X^T \hat{V}^{-1} Y \tag{3.15}$$

The covariance matrix of these estimates is

$$\text{cov}(\hat{\beta}) = (X^T \hat{V}^{-1} X)^{-1} X^T \hat{V}^{-1} \{\text{cov}(Y)\} \hat{V}^{-1} X (X^T \hat{V}^{-1} X)^{-1} \tag{3.16}$$

where $\text{cov}(Y) = V$ and is unknown because we cannot rely upon the usual normality assumption when estimating V (Appendix 2.1) or because the random part of the model is misspecified. The usual procedure would be to substitute the estimated $\hat{V}$, but this will generally lead to standard errors which are too small. A robust estimator is obtained by replacing $\text{cov}(Y)$ by $\tilde{Y}\tilde{Y}^T$, namely the cross product matrix of the raw residuals, which is a consistent estimator of $\boldsymbol{V}$. This is done for each highest level block of $\boldsymbol{V}$ in order to satisfy the block diagonality structure of the model**.** This estimator is a generalisation of the estimator given by Royall (1986) for a single level model which uses only the diagonal elements of $\tilde{Y}\tilde{Y}^T$. It is often known as a 'sandwich' estimator from the form of the right hand side in (3.16).

For the random parameters an analogous result holds. It is also possible to derive robust estimators for residuals, but these generally are not useful because the estimate for each residual corresponding to a higher level unit uses the corresponding value of $\tilde{Y}\tilde{Y}^T$ and this can give very unstable estimates.

We now apply (3.16) to the abortion data analyses and Table 3.5 shows the result for Model A of Table 3.4 and an OLS analysis. The major change is in the estimate of the standard error for the level 1 variance, with only moderate changes for the fixed parameters.

Another approach to providing robust standard errors is to use jacknifing (Miller, 1974). Thus, if we wished to calculate the standard error for a level 2 variance in a model with m level 2 units, the jacknife procedure would involve recomputing the variance for m subsamples, each one formed by omitting one level 2 unit, and using the set of these to form the standard error estimate. The procedure also gives a revised estimate of the parameter itself. Longford (1993, Chapter 6) gives an example in the analysis of a complex matrix sample design and suggests that there may be often a considerable loss of efficiency using the jacknife method, and it is also computationally intensive. A more flexible method is that of bootstrapping which we now describe.

3.6 The bootstrap

For a general introduction to bootstrapping see Efron and Gong (1983) and Laird and Louis (1987; 1989). Davison and Hinkley (1997) give more extensive discussions, especially in the context of a multilevel model. Three general bootstrapping approaches are available; a fully nonparametric bootstrap, a semiparametric or 'residuals' bootstrap and a fully parametric bootstrap.

3.6.1 The fully nonparametric bootstrap

The nonparametric bootstrap procedure for a single level model involves simple random resampling with replacement of the (level 1) units to generate a single bootstrap sample. The model parameter estimates are then re-estimated for this sample. This procedure is repeated a large number (N) of times yielding N sets of parameter estimates which are then treated as a simple random sample and used to derive standard errors or confidence intervals for parameters and also for bias correction. For a multilevel model, however, such a procedure is inadequate since it assumes identically distributed responses. An alternative that does provide suitable estimates is to resample only the higher level units, but if the number of such units is small this will not be very efficient. Nevertheless where the number of highest level units is large, for example typically in a repeated measures model, this procedure is useful.

No other procedure, for example, for a 2-level model sampling at level 2 and then at level 1 within level 2 units, seems able to preserve the original correlation structure since the bootstrap resampling procedure assumes independent responses for the level 1 units, and this is not the case for a multilevel structure. To preserve the correlation structure we should sample the full set of (correlated) level 1 units within each level 2 unit, that is, simply sample the higher level units.

3.6.2 The fully parametric bootstrap

The fully parametric bootstrap utilises the distributional assumptions of the model in order to generate simulated values which are used to estimate bootstrap sets of parameters. Consider the simple 2-level model assuming normality

$$y_{ij} = (X\beta)_{ij} + u_j + e_{ij}, \quad u_j \sim N(0, \sigma_u^2), \quad e_{ij} \sim N(0, \sigma_e^2)$$

To generate a bootstrap sample we select at random from $N(0, \sigma_u^2)$ a set of level 2 values u_j^* and for each level 2 unit a set of e_{ij}^* from $N(0, \sigma_e^2)$. These are added to $(X\beta)_{ij}$ to generate a set of pseudo values y_{ij}^* which are then treated as a set of responses from which a new set of bootstrap parameter values, $\hat{\beta}^*$, $\hat{\sigma}_u^{*^2}$, $\hat{\sigma}_e^{*^2}$ is obtained.

Once the set of bootstrap parameter values is available we can use these to estimate the parameter covariance matrices or standard errors using the usual procedures. Confidence intervals for the original parameter estimates or functions of them can then be constructed from these. Alternatively we can construct intervals nonparametrically from the percentiles of the set of empirical bootstrap values and where

the median value for a parameter or function of parameters deviates substantially from the original parameter estimate a bias correction procedure should be used. This involves smoothing the bootstrap distribution using, say, a standard normal distribution. We first estimate z_0 which is the standard normal score corresponding to the percentile position of the original parameter estimate. Writing $z^{(1-\alpha)}$, $z^{(\alpha)}$ for the standard normal deviates corresponding to the required (symmetric) percentiles (for example, 5 % and 95 %) we transform back to the bootstrap distribution from the standard normal distribution values

$$2z_0 + z^{(1-\alpha)}, \qquad 2z_0 + z^{(\alpha)}$$

Efron (1988) discusses this and a further correction based on skewness to improve accuracy.

If we wish to obtain bootstrap estimates for estimated level 2 residuals then for each bootstrap sample we also estimate the residuals, $\hat{u}_j^*$. To estimate the 'comparative' variance of the residuals for each level 2 unit we can deviate these from their mean to obtain their variance, or covariance matrix where there are several random coefficients. They can also be used to construct nonparametric confidence intervals as above.

Table 3.6 gives parametric bootstrap estimates of standard errors and a central 90 % confidence interval based upon a normality assumption and also a nonparametric estimation from 1000 bootstrap samples for the model of Table 3.5.

The bootstrap standard errors agree quite well with the model-based ones, except for the level 3 variance. This parameter is based upon only 54 level 3 units as opposed to 264 level 2 and 1056 level 1 units. This is reflected also in the bootstrap confidence intervals where the nonparametric intervals are fairly close to the normal theory ones except for the level 3 variance. In general, despite the computational overhead, bootstrap intervals will be desirable where effective sample sizes are small, especially for the random parameters. Where distributions are markedly nonnormal the nonparametric intervals are to be preferred, and 1000 is often regarded as a minimum.

3.6.3 The iterated parametric bootstrap and bias correction

We remarked in Section 2.4 that, in small samples, RIGLS (REML) estimation may be preferred since it produces unbiased estimates for the random parameters. Table 3.7 shows the comparison of the two estimates with a simple variance components model for the data of Table 3.1.

Since the number of level 2 units is moderately large the bias in the IGLS variance estimates is small, about 3 %. We can avoid this bias by carrying out a RIGLS estimation but we can also use the bootstrap as a bias correction tool, and we do this here for illustration only.

Assume we have carried out an IGLS estimation and obtained the results in column 1 of Table 3.6.

Table 3.6 Bootstrap standard errors and 90 % confidence intervals for Model A in Table 3.4. Bootstrap uses 1000 simulated datasets.

Parameter	Model-based s.e.	Bootstrap s.e.	normal C.I.	Nonparametric Adjusted C.I.
Fixed				
Religion:				
Roman Catholic	0.176	0.173	(−1.084, −0.516))	(−1.128, −0.532)
Protestant	0.098	0.100	(−0.429, −0.101)	(−0.420, −0.106)
Other	0.127	0.132	(−0.846, −0.414)	(−0.805, −0.377)
Year:				
1984	0.048	0.048	(−0.365, −0.209)	(−0.374, −0.216)
1985	0.048	0.047	(−0.140, 0.014)	(−0.141, 0.012)
1986	0.048	0.048	(−0.015, 0.141)	(−0.019, 0.141)
Random				
Level 3				
σ_v^2	0.030	0.022	([0], 0.066)	(0, 0.080)
Level 2				
σ_u^2	0.043	0.041	(0.302, 0.436)	(0.308, 0.438)
Level 1				
σ_{e0}^2	0.016	0.015	(0.284, 0.334)	(0.288, 0.336)

Table 3.7 JSP mathematics data. Normal score transform of 11-year score; original 8-year score centred about mean.

Parameter	IGLS Estimate (s.e.)	RIGLS Estimate (s.e.)	Bootstrap (1000) Estimate (s.e.)
Fixed			
Constant	0.137	0.137	0.136
8-year score	0.095 (0.004)	0.095 (0.004)	0.095 (0.004)
Gender (boys-girls)	−0.044 (0.050)	−0.044 (0.050)	−0.044 (0.049)
Social class (non-manual-manual)	−0.153 (0.057)	0.154 (0.057)	−0.152 (0.057)
Random			
Level 2			
σ_{u0}^2	0.0808 (0.0234)	0.0833 (0.0239)	0.0780 (0.0230)
Level 1			
σ_{e0}^2	0.4234 (0.0229)	0.4253 (0.0230)	0.4213 (0.0224)

We have

$$\begin{aligned} y_{ij} &= (X\beta)_{ij} + u_j + e_{ij} \\ u &\sim N(0, \sigma_u^2),\ e \sim N(0, \sigma_e^2) \end{aligned} \tag{3.17}$$

with estimates for the parameters $\hat{\theta}^T = (\hat{\beta}, \hat{\sigma}_u^2, \hat{\sigma}_e^2)$.

We now carry out a bootstrap and obtain the mean of the bootstrap parameter estimates $\bar{\theta}$, say. If, after a large number of bootstraps, $\bar{\theta}$ is different from $\hat{\theta}$ then the difference is an estimate of the bias of the estimation procedure we have used. In other words, assuming that the original estimates are in fact the 'true' model parameters, simulating from these and estimating parameters on average gives different parameter estimates from the original implying that our estimation procedure is biased. The estimate of the bias is simply $\bar{\theta} - \hat{\theta}$. A bias-corrected estimate is $2\hat{\theta} - \bar{\theta}$. We see from the third column of Table 3.7 that, using a bootstrap sample of 1000 we estimate a downward bias for the level 2 variance of 0.0808 – 0.0780 = 0.0028 and similarly for the level 1 variance a downward bias of 0.0021. The bias corrected estimates are therefore 0.0836 and 0.4255 respectively, which are, as expected, close to the RIGLS estimates. If we wish to calculate quantiles etc. from the set of bootstrap replicates then we should apply the bias correction to these estimates also.

In many cases, the bias associated with an estimation procedure is a function of the true parameters. Since we are simulating from the estimated parameters our bias estimate will itself generally be biased, and we can adopt an 'iterative bias correction'. This works as follows.

Carry out a bootstrap replicate set and obtain a bias-corrected vector of parameters, say $\hat{\theta}_1$. This will still, generally, be biased but less so than the original. Carry out a second bootstrap replicate set using the parameters $\hat{\theta}_1$ instead of the original estimates and calculate a new bias corrected estimate $\hat{\theta}_2$. While, generally, still biased, the bias will tend to be less than the previous bias. Repeat this using the new bias-corrected estimates until the sequence of estimates converges. In some cases convergence may not occur, but if it does we will obtain unbiased estimates (Kuk, 1995). This procedure is most useful when applied with discrete response models based upon quasilikelihood estimation, where in certain circumstances large biases are produced. A detailed description of the procedure for such models is given by Goldstein and Rasbash (1996) and we will use this in an example in Chapter 4.

3.6.4 The residuals bootstrap

One potential drawback to the fully parametric bootstrap is that it relies upon the (normal) distribution assumption for the residuals. In single level linear models a residuals (semi-parametric) bootstrap can be implemented by fitting the model, calculating the empirical residuals by subtracting the predicted response from the observed and then for each bootstrap iteration sampling from these residuals with replacement. In a multilevel model, however, the situation is more complicated. Consider first a 'crude' residuals bootstrap, as follows, for Model (3.17):

1. Estimate residuals ($\hat{u}_j$, $\hat{e}_{ij}$).
2. Sample with replacement m level 2 residuals and N level 1 residuals, add these to the fixed part estimate to generate a new set of Y values.
3. Fit the model to these new data and obtain the parameter estimates.
4. Repeat steps 2 and 3, say 1000 times.

Such a procedure leads to biases because the residuals are shrunken and the estimates across levels are correlated, negatively in the present case, so independent resampling is inappropriate. We require, therefore, both to 'reflate' residuals and create independence before resampling.

Consider first just reflating the residuals separately at each level. We illustrate with the 2-level model

$$\begin{aligned} y_{ij} &= (X\beta)_{ij} + (ZU)_j + e_{ij} \\ U^T &= \{U_0, U_1 \ldots .\} \end{aligned} \tag{3.18}$$

Having fitted the model, we estimate the residuals for each level 2 unit j

$$\{\hat{u}_{0j}, \hat{u}_{1j} \ldots .\},\ \hat{e}$$

For convenience, we shall illustrate the procedure using the level 2 residuals, but analogous operations can be carried out at all levels. Write the empirical covariance matrix of the estimated residuals at level 2 in (3.18) as

$$S = \frac{\hat{U}\hat{U}^T}{m} \tag{3.19}$$

and denote the corresponding model estimated covariance matrix of the random coefficients at level 2 as R. The empirical covariance matrix is estimated using the number of level 2 units, *m,* as divisor rather than *m-1.* We assume that the estimated residuals have been centred, although centring will only affect the overall intercept value. We also note that no account is taken of the relative sizes of the level 2 units and we could consider a weighted form of (3.19)

$$S = \frac{\hat{U}W\hat{U}^T}{N} \tag{3.20}$$

where W is the (m x m) diagonal matrix with unit sizes on the diagonal. This is equivalent to defining new residual variables

$$U'_{hj} = U_{hj}\sqrt{mn_j/N} \tag{3.21}$$

and using these with (3.19).

For (3.19) we now seek a transformation of the residuals of the form

$$\hat{U}^* = \hat{U}A$$

where A is an upper triangular matrix of order equal to the number of random coefficients at level 2, and such that

$$\hat{U}^{*T}\hat{U}^*/m = A\hat{U}^T\hat{U}A = A^T SA = R \tag{3.22}$$

so that these new residuals have the required model covariance matrix, and we now can sample sets of residuals with replacement from $\hat{U}^*$. This will be done at every level of the model, with sampling being independent across levels, thus retaining the independence assumption of the model. Having sampled a set of these residuals we add them to the fixed part of the model, along with correspondingly sampled level 1 residuals to obtain the new set of responses.

We can form A as follows: write the Cholesky decomposition of S, in terms of a lower triangular matrix as
$S = L_S L_S^T$ and the Cholesky decomposition of R as $R = L_R L_R^T$. We have

$$L_R L_S^{-1}\hat{U}^T\hat{U}(L_R L_S^{-1})^T = L_R L_S^{-1} S(L_S^{-1})^T (L_R)^T = L_R (L_R)^T = R$$

and the required matrix is therefore $A = (L_R L_S^{-1})^T$.

Carpenter *et al.* (1999) use this procedure, unweighted, to demonstrate the improved (confidence interval based) coverage probabilities compared to the parametric bootstrap when the level 1 residuals have a chi-squared distribution rather than a normal. Further work confirms this but also suggests that the procedure may underestimate coverage for certain departures from an assumed normal distribution. (Carpenter *et al.*, 2003). If we apply this procedure to the data in Table 3.6 we obtain estimates very similar to those using the fully parametric bootstrap. This procedure can also be iterated in the same way as the parametric bootstrap.

The above reflating procedure takes no account of dependencies across levels. If we now write

$$Q^T = \{U_0, U_1 \ldots, e\}, Q^T \text{is } (N \times [p+1]) \tag{3.23}$$

where p is the number of level 2 random effects and Q^T is the length of the data matrix, then analogously to (3.19) we form

$$S = \frac{\hat{Q}\hat{Q}^T}{N} \tag{3.24}$$

and proceed as before with computing transformed residual sets for the resampling bootstrap. The R matrix has the form

$$\begin{pmatrix} \Omega_u & 0 \\ 0 & \sigma_e^2 \end{pmatrix}$$

so that the Cholesky decomposition is formed in the same way as the separate ones above. The S matrix, however, does not have this form since there are cross product terms for the level 2 and level 1 residuals of the form

$$\frac{1}{N} \sum_j \hat{u}_{hj} \sum_i \hat{e}_{ij} \tag{3.25}$$

We note that the Q vectors are still uncorrelated across levels so that we can sample them separately at each level and maintain the model independence assumption as before. We also note that if we ignore the terms (3.25) we obtain the weighted procedure given by (3.20).

3.7 Aggregate level analyses

As discussed in Chapter 1, there are sometimes occasions when the only data available for analysis have already been aggregated to a higher level. For example, we may have information on student achievement in terms only of the mean achievement for each school, or information on utilisation of health services in terms only of the total number of episodes for each administrative area. We now examine the possibilities for carrying out analyses with aggregate level data and explore how far these can provide information about the parameters of a more disaggregated model.

Consider the simple model for the Junior School Project data, with the 11-year mathematics test score as response and the earlier mathematics score as a covariate

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + u_j + e_{ij} \tag{3.26}$$

Suppose that we now aggregate to the school level by averaging over all pupils in each school to obtain

$$y_{.j} = \beta_0 + \beta_1 x_{.j} + u_j + e_{.j} \tag{3.27}$$

Fitting this as a single level model, the level 1 variance is $\sigma_u^2 + n_j^{-1}\sigma_e^2$ and we can fit the model by specifying two explanatory variables for the random part, namely

$$z_0 = 1,\ z_{1j} = n_j^{-0.5}$$

with random coefficients e_{0j}, e_{1j} each having a variance with zero covariance. In many surveys the same number of level 1 units will be sampled from each level 2

Table 3.8 School level analysis of JSP data.

Parameter	Estimate (s.e.) Model A	Estimate (s.e.) Model B	Estimate (s.e.) Model C
Fixed			
Constant	0.18	0.16	0.16
8-year score	0.091 (0.019)	0.092 (0.020)	0.094 (0.021)
Gender (proportion boys)	−0.34 (0.30)	−0.31 (0.30)	−0.29 (0.29)
S. Class (proportion N.M.)	0.00 (0.20)	0.00 (0.28)	−0.01 (0.27)
Random			
σ^2_{u0}	0.11 (0.021)	0.11 (0.040)	0.08 (0.024)
σ^2_{e0}		0.08 (0.37)	
σ_{u01}			0.00 (0.01)
σ^2_{u1}			0.004 (0.004)
−2 (loglikelihood)	31.33	31.28	29.44

unit, in which case a single explanatory variable z_0 will suffice, but this does not then allow us to estimate the level 1 and level 2 variances separately. The other problem with such an analysis is that the estimates will be inefficient compared with those from a 2-level model based on individual student data.

Model A in Table 3.8 gives the results of an analysis using just the single explanatory variable z_0 and Model B additionally uses z_{1j} and so is equivalent to a single level weighted regression model. In both analyses we have included the proportion of non-manual students and the proportion of girls as explanatory variables, that is the average values of the corresponding (0.1) dummy variables.

In comparison with the analyses in Table 3.7, while the coefficient of the 8-year maths score changes only slightly, those for gender and social class change markedly. We also see how the standard errors are substantially greater. In fact, although the number of students per school varies between 3 and 49, the inclusion of z_{1j} has little effect.

For these data we know that the slope of the 8-year score is random across schools. In this case Model (3.27) becomes

$$y_{.j} = \beta_0 + \beta_1 x_{.j} + u_{0j} + u_{1j} x_{.j} + e_{.j} \tag{3.28}$$

and we obtain the additional contributions to the variance of the aggregated level 2 units

$$\sigma^2_{u1} x^2_{.j}, \quad 2\sigma_{u01} x_{.j}$$

Model C in Table 3.8 shows the results of fitting this model, giving a poor estimate of the random coefficient variance, and unlike Model B it is not possible to estimate a separate level 1 variance because of the small number of units in the analysis.

If there is complex variation at level 1, such as we fitted in Table 3.2, then for such an explanatory variable, say z_{2ij}, we would obtain the further contributions to the variance for the aggregated model for unit j

$$2\sigma_{e02}z_{2.j}/n_j, \quad \sigma^2_{e2}\sum_i z^2_{2ij}/n^2_j$$

The first of these terms can be fitted as a covariance and the second as a variance, by defining appropriate explanatory variables. In the present case the data are not extensive enough to allow us to fit these additional variables. We also note that the values of the squared explanatory variables in the second of these expressions will often not be available for aggregated data.

If we have an initial 3-level model, and data are aggregated to level 2, we need to specify properly the level 2 random variation resulting from the aggregation process. Failure to do this, may allow us to fit random variation at level 3, but any interpretation of this may be problematic because it may have arisen solely as a result of misspecifying the variation at level 2. For example, if we have an explanatory variable which is strongly correlated with the size of the level 2 units, and we fail to include a random coefficient for z_{1j} at level 2, we may well be able to fit a random coefficient for it at level 3, but the usual interpretation of such a coefficient would be incorrect.

We now look at what happens to the fixed part coefficients when aggregation takes place and we have already seen that the values of the coefficients for gender and social class change. Consider the 2-level model

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 x_{.j} + u_j + e_{ij} \tag{3.29}$$

where the coefficient for $x_{.j}$ in the aggregated model is now $\beta_1 + \beta_2$. We saw in Table 3.1 that the coefficient for the school mean 8-year score was very small, so that we would expect the coefficient for this in the aggregated model to be similar, which Table 3.8 confirms. For gender and social class the coefficients of the corresponding aggregated variables from a 2-level analysis are respectively –0.06 and –0.09, which when added to the (non-aggregated) coefficients for gender and social class give values of –0.09 and –0.06 respectively. These are rather different from those in Table 3.1, but the standard errors are very large. Where there is a contextual or compositional effect, whether through the mean aggregated value, or some other statistic derived from the student level distribution, then an aggregated analysis will not allow us to obtain separate estimates for the individual and compositional coefficients.

3.7.1 Inferences about residuals from aggregate level analyses

The use of aggregate level analysis can be particularly misleading for residual estimates. Woodhouse and Goldstein (1989) show how this can occur for educational

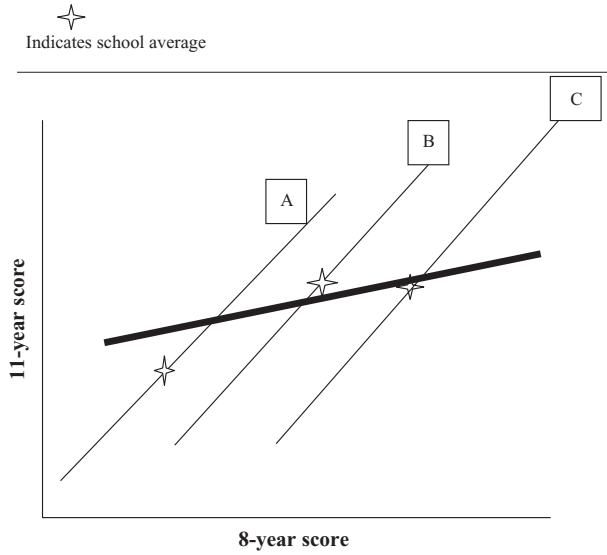


Figure 3.7 Aggregate versus individual level model inferences; hypothetical illustration for three schools.

data aggregated to the level of Education Authority and Figure 3.7 illustrates a situation where the use of aggregate data at face value provides incorrect inferences. For the aggregate analysis the school ordering for given 8-year score is: $A<C<B$. For the pupil level analysis the school ordering is: $A>B>C$, which gives a quite different conclusion. In general, the only circumstance where we can be certain that aggregate and individual level inferences are the same is where the explanatory variable distribution (that of the 8-year score here) is the same for each higher level unit (school).

In summary, we have seen that it is sometimes possible to model aggregated data, but this has to be carried out with care, and any interpretations will be constrained by the nature of the true, underlying, nonaggregated model. In addition, the precisions of the estimates obtained from an aggregated analysis will generally be much lower than those obtained from a full multilevel analysis. A discussion of the aggregation issue can also be found in Aitkin and Longford (1986).

3.8 Meta analysis

The term 'meta analysis' (Hedges and Olkin, 1985) refers to the pooling of results of separate studies, all of which are concerned with the same research hypothesis. The aim is to achieve greater accuracy than that obtainable from a single study and also to allow the investigation of factors responsible for between-study variation. Each study typically provides an estimate for an 'effect', for example, a group difference,

for a 'common' response and the original data are often unavailable for analysis. In general, the response measures used will vary, and care is needed in interpreting them as meaning the same thing. Furthermore, the scales of measurement may differ, so that the effect is usually standardised using a suitable within-study estimate of between-unit standard deviation. If the study result derives from a multilevel model, then this estimate will be based on the level 1 variance, or where this is complex on an estimate pooled over the effect groups being compared. It is important that comparable estimates are used from each study. This implies that the specification of the level 1 units is comparable and that the sources of higher level variation are properly identified. For example, for a set of studies comparing teaching methods using a number of schools, the within-school between-student variation would be appropriate for standardisation, which implies ideally that the studies concerned should provide estimates of this using suitable multilevel techniques. We consider the case where only a single effect is of interest, but the generalisation to the multivariate case is straightforward (see Chapter 6).

In the standard case, for the *j-th* study we define the standardised effect d_j where this is a dimensionless quantity. It may, for example, be a correlation coefficient, a standardised regression coefficient, a group difference, or a weighted group difference. We require an estimate of the variance of d_j, say σ_j^2, and more generally we require the variance of a dimensionless function having the general form

$$\sum_h w_{hj}\hat{\beta}_{hj}/\hat{\sigma}_{ej} \tag{3.30}$$

where the $\hat{\beta}_{hj}$ are parameter estimates from the j-th study. A pooled estimate can now be derived using these standardised effects and their variance estimates. For moderately large numbers of level 1 units, we can ignore the variation in the estimate of the level 1 standard deviation ($\hat{\sigma}_{ej}$) and calculate the variance of the numerator of (3.30) using the estimated covariance matrix of the $\hat{\beta}_{hj}$.

We may have several studies each of which may provide information on treatment effects at different levels of aggregation. We can formulate a general class of meta analysis models by considering a simple 2-level structure. Assume that we have a collection of studies, each concerned with the comparison of several 'treatments'. These treatments may be distinct categories (represented by dummy variables) or may be effects represented by regression coefficients or a mixture of the two kinds. The basic models we shall develop are 'variance component' models but we will also illustrate the use of a random coefficient model, and the variance heterogeneity (complex variation) case can also be incorporated.

For the i-th subject in the j-th study who received the h-th treatment, we can write a basic underlying model for outcome y_{hij} as

$$\begin{aligned} y_{hij} &= (X\beta)_{ij} + \alpha_h t_{hij} + u_{hj} + e_{hij} \\ h &= 1, \ldots, H; \quad j = 1, \ldots, J; \quad i = 1, \ldots, n_{hj} \\ u_{hj} &\sim N(0, \sigma_{hu}^2); \quad e_{hij} \sim N(0, \sigma_{he}^2) \end{aligned} \tag{3.31}$$

where $(X\beta)_{ij}$ is a linear function of covariates for the i-th subject in the j-th study, $u_{hj}$is the random effect of the h-th treatment for the j-th study and e_{hij} is the random residual of the h-th treatment for subject i in study j. The term t_{hij} is a dummy treatment variable (contrasted against a suitable base category) and α_h is the treatment contrast of interest. If the treatment dummy variables are replaced by a continuous variable t_{ij} then (3.31) becomes

$$\begin{aligned} y_{ij} &= (X\beta)_{ij} + \alpha t_{ij} + u_j + e_{ij} \\ j &= 1, \ldots, J; \quad i = 1, \ldots, n_j \\ u_j &\sim N(0, \sigma_u^2); \quad e_{ij} \sim N(0, \sigma_e^2) \end{aligned}$$

It is also clearly possible to allow the variances within and between studies to be different for each treatment or to vary with the value of a continuous treatment variable, leading to complex variance structures. We can also introduce covariates where data are available and appropriate, and interactions between treatments and covariates. For example, a particular treatment contrast may differ according to covariate values. We may also relax the normality assumption of the level 1 residuals, for example if fitting a generalised linear multilevel model (see Turner *et al.*, 2000 for an example). We can also model the case where different sets of treatments occur in different studies. For example, a placebo may be compared with a different treatment in each study. Such models are sometimes known as network meta analyses (Lumley, 2002).

We can extend (3.31) to the multivariate case, using the techniques of Chapter 6, where several responses are being investigated simultaneously and interest centres on their interrelationships.

3.8.1 Aggregate and mixed level analysis

Consider now the case where (3.31) is the underlying model but we only have data by treatment group at the study level. Aggregating to this level, as in Section 3.7, we write the mean response as

$$y_{h.j} = (X\beta)_{.j} + \alpha_h t_{h.j} + u_{hj} + e_{h.j} \tag{3.32}$$

where the '.' notation denotes the mean for study j. A difficulty may arise with the first term in (3.32) since this implies that the mean of the covariate function $(X\beta)_{ij}$ for each study is available.

The corresponding model for the case of a continuous treatment variable is

$$y_{.j} = (X\beta)_{.j} + \alpha t_{.j} + u_j + e_{.j}$$

Consider the special case of two treatments, $h = 1,2$. We collapse (3.32) and, using an obvious notation, rewrite it to give

$$\begin{aligned} y'_{.j} &= y_{1.j} - y_{2.j} = \alpha + u'_j + e'_{.j} \\ \alpha &= \alpha_1 - \alpha_2 \end{aligned} \tag{3.33}$$

This impliesvar$(u'_j) = \text{var}(u_{1j}) + \text{var}(u_{2j}) - 2\text{cov}(u_{1j}, u_{2j})$. We can also combine (3.31) and (3.32) into a single model for the case where some aggregated responses are in terms of separate treatment groups and some are in terms of contrasts of groups.

3.8.2 Definin origin and scale

When combining data from aggregate level studies it is necessary to ensure that the response variable scales are the same and that there is a common origin. In traditional two-treatment meta analyses the treatment difference is divided by a suitable (pooled) within-treatment standard deviation. In our general model, likewise, the response variable in each study can be scaled by dividing it by an estimate of the level 1 standard deviation. Where individual data are available we may use an estimate of the level 1 standard deviation from a preliminary analysis and for aggregate data we may derive this from reported summary information, if this is available.

In situations where the same response variable is used in each study, and scaling has been carried out, we can apply (3.32) and (3.33) directly. In many cases, however, different response variables are used. For example, in class size studies different reading tests may be used. In this case we would not generally expect the means for corresponding treatments to be identical. One procedure for dealing with this is to choose one treatment as a reference treatment (or control) and in each study subtract its mean from the values of the other treatments and work *with these differences*. This is the standard approach in two-treatment studies and such differences can either be fixed effects or vary randomly across studies.

3.8.3 An example: meta analysis of class size data

Goldstein *et al*. (2000b) consider the meta analysis of nine studies of the effects of class size on the learning of young children. Eight of these studies report only aggregate level data and one (STAR) provides individual student data. The studies were selected, from thousands of reported studies, as being those that satisfied stringent quality criteria for design and reporting of data. Applying the above models, after suitable scaling, a joint analysis gives the following results.

In the analysis using only the STAR individual level data, we can distinguish three levels: individual, class and school. The response has been normalised so that the level 1 variance is fixed at 1.0 and we see that there is an overall negative relationship with class size whereby there is a gain of about a quarter of a standard deviation for a reduction in class size of 10 children. There is also reasonably large variation between-classes, within-schools, and between-schools in the magnitude of this effect. The STAR study was a randomised controlled trial and its design

Table 3.9 Combined analysis of class size study data. Reading score response at end of year 1 (post kindergarten).

	STAR data only	Combined data
Fixed		
α_0	0.078	0.184
α_1 (class size, linear)	−0.024 (0.006)	−0.022 (0.007)
α_2 (pre-test)	0.907 (0.018)	0.907 (0.018)
Random		
Level 4 (between-study)		
σ^2_{w0}		0.038 (0.020)
σ_{w01}		−0.004 (0.002)
σ^2_{w1}		0.0004 (0.0002)
Level 3 (between-school)		
σ^2_{v0}	0.305 (0.064)	0.305 (0.064)
σ_{v01}	0.00014 (0.004)	0.00012 (0.004)
σ^2_{v1}	0.0006 (0.0006)	0.0006 (0.0006)
Level 2 (between-class)		
σ^2_{u0}	0.139 (0.023)	0.138 (0.023)
Level 1 (between-student)		
σ^2_e	1.000 (0.023)	1.000 (0.023)
-2 log likelihood	11996.5	11948.3

raises some particular issues of interpretation which are discussed by Goldstein and Blatchford (1998).

In the combined analysis the eight non-randomised, but longitudinal studies reporting at aggregate level are combined with the STAR study. This allows us to estimate between-study variation at level 4 and there is some evidence for the relationship varying across studies, although the combined mean estimate differs little from that for the STAR study alone.

We show in Chapter 6 how to extend single response models to multivariate response models and this can be applied to the meta analyses discussed above. A discussion of such models can be found in Riley (2009).

3.8.4 Practical issues in meta analysis

There are a number of practical problems with meta analysis studies. One of these is where the sample of studies used is subject to systematic bias. This can occur, for example if some studies do not provide sufficient data to estimate a standardised difference and they are a special group. Another common problem arises where the analysis is based upon published studies and those studies which found 'nonstatistically significant' results tend to remain unpublished. This implies that the distribution

of results is censored with the smaller ones tending to be missing, a situation known as the publication bias effect. In this connection, Higgins *et al*. (2009) discuss Bayesian models for meta analysis and emphasise the need to study the variability of effects across studies. Turner *et al*. (2009) discuss practical ways of correcting for bias in meta analyses and suggest models for doing this. Welton *et al*. (2009) also discuss this issue and present a model where the bias is assumed to have a distribution across studies.

3.9 Design issues

When designing a study where the multilevel nested structure of a population is to be modelled, the allocation of level 1 units among level 2 units and the allocation of these among level 3 units etc. will clearly affect the precision of the resulting estimates of both the fixed and random parameters. The situation becomes more complex, for example when there are cross classifications and where there are several random coefficients. There are generally differential costs associated with sampling more level 1 units within an existing level 2 unit as opposed to selecting further level 1 units in a new level 2 unit.

Some approximations for studying the standard errors of the fixed coefficients have been derived by Snijders and Bosker (1993) in the case of a simple 2-level variance components model. They are concerned with students sampled within schools and assume that the cost of selecting a student in a new school is a fixed constant times the cost of selecting a student in an already selected school. They also assume that there is a minimum of 11 students per school. They tend to find that, where this constant is greater than 1 and the total number of students to be sampled is fixed, the sample of schools should be as large as possible, although this will not necessarily be true for all the coefficients of interest.

Moerbeek *et al*. (2000) consider a 3-level model where interest focuses on the estimation of a treatment effect so that the aim is to minimise the variance of the relevant fixed coefficient. They consider the balanced case and show analytically how differential allocations across units can minimize overall cost. In particular they consider the allocation of different treatments to level 1 units within level 2 or level 3 units versus allocation at the higher unit level. Moerbeek *et al*. (2001) extend their analysis to the 2-level binary response model (see Chapter 4) and show how similar results can be obtained using MQL estimation. They show, in the case of a particular simulation study that design decisions based upon MQL are similar to those when either PQL or maximum likelihood estimation is used. They also discuss practical and ethical issues associated with allocation within higher level units, especially where these are for example schools, where 'contamination' and lack of independence can arise (see also Goldstein and Blatchford, 1998, for a discussion in the context of educational interventions).

In practice, balanced designs are a special case and interest focuses on both fixed and random parameters. Analytical results are generally not available for unbalanced designs, but a guide to design efficiency can be obtained by simulating the effect of

different design strategies and studying the resulting characteristics of the parameter estimates, such as their mean squared errors. This will be time consuming however, since for each design a number of simulated samples will be required. Mok (1995) carries out such an analysis for a 2-level random coefficient model and makes tentative recommendations based upon the intra unit correlation and numbers of level 1 and level 2 units.

Cohen (1998) considers the optimality (efficiency) with respect to various fixed and random effects in a 2-level balanced model as a function of the intra-school correlation and shows that optimality criteria can vary considerably with the choice of parameter. Moerbeek and Wong (2002) extend these results by considering the robustness to misspecification of the intra unit correlation and also consider ways of combining optimal choices across parameters. They consider the case where weights are chosen to minimise a linear function of parameter variances. It is also possible to consider an optimality criterion based, for example, upon the determinant of the covariance matrix of the fixed or random parameters.

More recently, Browne *et al*. (2009) have produced a general package that will carry out power calculations for a variety of multilevel models with balanced or unbalanced designs, including cross-classified models. This is available as a free-standing package, or as a set of macros that can be used within MLwiN or R. Raudenbush (2009) has also developed a free-standing package.

4

Multilevel models for discrete response data

4.1 Generalised linear models

All the models of previous chapters have assumed that the response variable is continuously, and in particular normally, distributed. We now look at data where the response is essentially a count of events. This count may be the number of times an event occurs out of a fixed number of 'trials', in which case we usually deal with the resulting proportion as response: an example would be the proportion of deaths in a population, classified by age. We may also have a vector of counts representing the numbers of events of different kinds which occur out of a total number of events: an example is given in Chapter 3, where we studied the number of responses to each ordered category of a question on abortion attitudes.

Statistical models for such data belong to the class of 'generalised linear models' (McCullagh and Nelder, 1989). A 2-level model can be written in the general form

$$\pi_{ij} = f(X\beta)_{ij} \tag{4.1}$$

where π_{ij} is the expected value of the response for the ij-th level 1 unit and f is a nonlinear function of the 'linear predictor' $(X\beta)_{ij}$. Note that we can allow random coefficients at level 2. The model is completed by specifying a distribution for the *observed* response $y_{ij}|\pi_{ij}$. Where the response is a proportion, this is typically taken to be binomial and where the response is a count taken to be Poisson. It remains for us to specify the nonlinear 'link' function f. Table 4.1 lists some of the standard choices, with logarithms chosen to base e. We see later how functions of these applied, for example, to ordered categories, can be specified.

Multilevel Statistical Models: 4th Edition Harvey Goldstein

Table 4.1 Some basic nonlinear link functions.

Response		Name
Proportion	$f^{-1} = \log\{(\pi)/(1-\pi)\}$	logit
Proportion	$f^{-1} = \log\{-\log(1-\pi)\}$	complementary log-log
Proportion	$\pi = \int_{-\infty}^{(X\beta)} \phi(t)dt$	Probit
Vector of proportions	$f^{-1} = \log(\pi_s/\pi_t)(s = 1, \ldots, t-1)$	multivariate logit
Vector of proportions	$f^{-1} =$ $\log\left\{\log\left(\frac{1}{t-1} + \frac{\pi^{(s)}}{\pi^{(t)}}\right)\right\} = (X\beta)$, $(s = 1, \ldots, t-1)$ $\pi^{(s)} = \left(e^{e^{(X\beta)^{(s)}}} - \frac{1}{t-1}\right)\left(\sum_{h=1}^{t-1} e^{e^{(X\beta)^{(h)}}}\right)^{-1}$	Multivariate complementary log-log
Count	$f^{-1} = \log(\pi)$	Log

We can also have the 'identity' function $f^{-1}(\pi) = \pi$, but this can create difficulties since it allows, in principle, predicted counts or proportions which are, respectively, less than zero or outside the range (0,1). Nevertheless, in many cases, using the identity function produces acceptable results which may differ little from those obtained with the nonlinear functions. In the following sections we consider each common type of model in turn with examples. We shall also be introducing new estimation procedures based upon maximum likelihood, MCMC, and quasilikelihood approximations.

4.2 Proportions as responses

Consider the 2-level variance components model with a single explanatory variable where the expected proportion is modelled using a logit link function

$$\pi_{ij} = \{1 + \exp(-[\beta_0 + \beta_1 x_{1ij} + u_{0j}])\}^{-1} \tag{4.2}$$

The observed responses y_{ij} are proportions with the standard assumption that they are binomially distributed

$$y_{ij} \sim Bin(n_{ij}, \pi_{ij}) \tag{4.3}$$

where n_{ij} is the denominator for the proportion. We also have

$$\text{var}(y_{ij}|\pi_{ij}) = \pi_{ij}(1-\pi_{ij})/n_{ij}$$

To emphasise the link with continuous responses and to introduce quasilikelihood estimation we now write the level 1 component as

$$y_{ij} = \pi_{ij} + e_{ij} z_{ij}, \quad z_{ij} = \sqrt{\pi_{ij}\left(1 - \pi_{ij}\right)/n_{ij}}, \quad \sigma_e^2 = 1 \tag{4.4}$$

Using this explanatory variable Z and constraining the level 1 variance associated with its random coefficient (e_{ij}) to be one, we obtain the required binomial variance in (4.4). When fitting a model we can also allow the level 1 variance to be estimated and by comparing the estimated variance with the value 1.0 obtain a test for 'extra binomial' variation. Such variation may arise in a number of ways.

For example, if we have omitted a level in the model, say ignored household clustering in a survey with one or more individuals sampled from a household, we would expect a greater than binomial variation at the individual level. Likewise, suppose the individuals and households were nested within areas and we chose to classify individuals, say by gender and three social class groups giving six cells in each area. If we treat these as the level 1 units so that the response is a proportion, then we no longer have a binomial variance, since these proportions are based upon the sum of separate binomial variables with differing probabilities. Here the variance for cell j within an area would have the form

$$[\pi_j(1 - \pi_j) - \sigma_1^2]/n_j$$

where n_j is the cell size. To fit such a model, we would specify an extra level 1 explanatory variable equal to $1/\sqrt{n_j}$ for the j-th cell, with 'variance' parameter at level 1 which is allowed to be negative (see Section 3.1). More generally, we can fit a model with an extra binomial parameter together with a further term such as above to give the following level 1 variance structure (omitting subscripts)

$$[\sigma_0^2 \pi\,(1 - \pi) + \sigma_1^2]/n$$

We do not, of course, know the true value π_{ij} or π_j so that at each iteration we use estimates based upon the current values of the parameters. Because we are using only the mean and variance of the binomial distribution to carry out the estimation, the estimation is known as 'quasilikelihood' (see Appendix 4.1 for details).

Another way of modelling such extra binomial variation, which has certain advantages, is to insert a 'pseudo level' above level 1. Thus, for individuals sampled within households, level 1 would be that of the individual and we would specify level 2 as that of the individuals also to give exactly 1 level 1 unit per level 2 unit. We specify binomial variation at level 1 and at level 2 we can now fit further random coefficients. For example, if we fit a random coefficient for the explanatory variable (with a 'variance' parameter which can be allowed to be negative) this is equivalent to specifying an extra level 1 variable $1/\sqrt{n_j}$ as above. In the above example where individuals are classified by gender and social class we can create a level 2 unit coinciding with each level 1 unit, fit binomial variation at level 1 and add level 2 variation which is

a function of gender and social class, for example an additive function with four parameters (see Chapter 3). We may wish to model the between-area variation of the cell proportions in terms of a simple variance term, rather as inversely proportional to n_j, in which case we would choose a simple dummy variable structure rather than explanatory variables proportional to $1/\sqrt{n_j}$. This 'pseudo level' procedure is rather similar to the way in which a meta analysis with known level 1 variation is modelled (Chapter 3).

In Appendix 4.1, we describe a linearisation procedure that allows us to use existing IGLS and RIGLS estimation for generalised linear models. Thus, for the model given by (4.2) and (4.4) a second order Taylor series expansion about the current values of the estimated residuals gives

$$\begin{aligned}\pi_{ij} &= f_{ij}(H_t) - X_{2ij}\beta_{2,t} f'_{ij}(H_t) + (Z_{2ij}\hat{u}_{2j}) f'_{ij}(H_t) \\ &\quad + X_{2ij}\beta_{2,t+1} f'_{ij}(H_t) + (Z_{2ij}u_{2j}) f'_{ij}(H_t) + Z_{2ij}(u_{2j} - \hat{u}_{2j})^2 f''_{ij}(H_t)/2 \qquad (4.5) \\ H_t &= X_{2ij}\beta_{2,t} + Z_{2ij}\hat{u}_{2j}\end{aligned}$$

where the terms on the right-hand side of the first line of (4.5) form an offset subtracted from the response at each iteration. We refer to estimates based upon (4.5) as PQL2 estimates; that is, predictive quasilikelihood with a second order approximation. If a first order approximation is used these are PQL1 estimates and if the expansion is about zero rather than the current residual estimates ($\hat{u}_{2j}$) these are known as marginal quasilikelihood estimates, MQL2 and MQL1 respectively.

In many applications, the MQL procedure will tend to underestimate the values of both the fixed and random parameters, especially where n_{ij} is small. Greater accuracy is obtainable if the second order approximation is used rather than the first order based upon the first term in the Taylor expansion. Also, when the sample size is small the unbiased (RIGLS, REML) procedure should be used. To illustrate the difference, Table 4.2 presents the results of simulating the following model, where the response is

Table 4.2 Mean values of 400 simulations. Empirical standard error in first bracket; mean of estimated standard errors in second bracket (IGLS).

	True $\sigma^2_{u0} = 0.5$		True $\sigma^2_{u0} = 1.0$	
Parameter	MQL first order	PQL second order	MQL first order	PQL second order
σ^2_{u0}	0.386(0.115) (0.130)	0.480(0.157) (0.152)	0.672(0.157) (0.188)	0.964(0.278) (0.255)
β_0	0.448(0.126) (0.129)	0.499(0.139) (0.138)	0.420(0.145) (0.149)	0.500(0.171) (0.172)
β_1	0.934(0.154) (0.147)	1.018(0.168) (0.154)	0.875(0.147) (0.145)	1.017(0.171) (0.158)

binary (0,1). The example assumes one moderate and one fairly large level 2 variance.

$$\begin{aligned}
&\text{logit}(\pi_{ij}) = \beta_0 + \beta_1 x_{ij} + u_{0j} \\
&y_{ij} \sim Bin(1, \pi_{ij}) \\
&\text{var}(u_{0j}) = 0.5, 1.0 \\
&\beta_0 = 0.5, \quad \beta_1 = 1.0
\end{aligned}$$

There are 50 level 2 units with 20 level 1 units in each level 2 unit. The following results are based upon 400 simulations of the above model for each variance value.

It is clear that the MQL first order model underestimates all the parameter values, whereas the second order PQL model produces estimates closer to the true values. The estimates given are based upon IGLS. In every case, convergence was achieved in less than 10 iterations. Very similar estimates for the fixed coefficients are obtained using RIGLS, and for the level 2 variances the PQL estimates become 0.498 and 0.996 respectively, which are closer to the true values. In addition, the averages of the standard errors given by both models are reasonably close to those calculated empirically from the replications. If we calculate 95 % confidence intervals for the parameters in the second order PQL model using the estimated standard errors and assuming normality then for the variance, we find that about 91 % of the intervals include the true value and for β_0 and for β_1 about 95 % do so. Hence, inferences about the true values would not be too misleading. The results of Table 4.2 are based upon a balanced dataset with equal numbers of level 1 units within each level 2 unit. In other cases, none of these procedures works satisfactorily. Thus, for example, when the average observed probability is very small (or very large), if many of the level 2 units have few level 1 units and there are very few level 2 units with large numbers of level 1 units, we will often find: that where the response is binary, there will be many level 2 units where the responses are all zero; that convergence with PQL2 or PQL1 often may not be possible; and that even where estimates are obtained, they may be substantially biased.

Rodriguez and Goldman (2001) discuss this issue and make recommendations. Where these quasilikelihood methods cannot be used, we may carry out a full maximum likelihood estimation (Appendix 4.2), use MCMC (Appendix 4.3) or carry out an iterated bootstrap (Appendix 4.4); these procedures are illustrated in the following example.

4.3 Examples

4.3.1 A study of contraceptive use

This dataset comes from the 1988 Bangladesh Fertility Survey (Huq and Cleland, 1990). It consists of 1934 women who are grouped in 60 districts; the response of interest is whether these women were using contraceptives at the time of the survey. Explanatory variables include age, the number of existing children and whether the district is urban or rural.

Table 4.3 Contraceptive use in Bangladesh. Standard errors in brackets.

Fixed parameter	A (MQL1)	B (PQL2)
Intercept	−1.59	−1.72
Age*	−0.025 (0.008)	−0.026 (0.008)
Urban	0.754 (0.162)	0.816 (0.170)
1 child	1.06 (0.16)	1.14 (0.16)
2 children	1.27 (0.17)	1.36 (0.18)
3+ children	1.26 (0.18)	1.36 (0.18)
Random parameter		
σ^2_{u0} (intercept)	0.334 (0.103)	0.396 (0.118)
σ_{u01}	−0.353 (0.142)	−0.414 (0.160)
σ^2_{u1} (urban)	0.596 (0.258)	0.685 (0.284)

*Age is centred at the mean age of 29.56. The base categories for other variables are 'rural' and 0 children. RIGLS estimation.

The MQL1 estimates are fairly close to the PQL2 estimates and in fact there are only 2 level 2 units (districts) with less than 10 women. There is a clear relationship with age, younger women being more likely to use contraceptives. Urban women are far more likely to use contraceptives, the odds ratio being $e^{0.82} = 2.3$. In terms of number of children, the greatest difference is in moving from no children to one child, an odds ratio of 3.1, with little change after 2 children. There is a relatively large variation in the between-district urban-rural difference. If we allow extra-binomial variation for the PQL2 model, the estimate of the multiplicative term is 0.96 with a standard error of 0.03; and if we fit an additive term we obtain a value of –0.09 with a standard error of 0.15. Neither of these estimates provides evidence for extra binomial variation. A large sample 'Wald' test (Chapter 2) gives $\chi^2 = 6.96$, $P < 0.05$.

Figure 4.1 shows the standardised residual plots for the intercept and coefficient of urban locality.

These plots produce approximately straight lines. Some care is needed with such plots when the number of level 1 units per higher level unit is small. The residual estimate is a linear function of binary responses and even where the underlying higher level distribution is normal, we will need a reasonably large number of these responses to approximate it adequately, especially with very small or very large probabilities.

Table 4.4 shows the parameter estimates for the model of Table 4.3, using MCMC, maximum likelihood and the iterated bootstrap (Appendix 4.4). The maximum likelihood estimates are obtained using simulated maximum likelihood with 400 simulated replicates. The MCMC estimation uses 5000 iterations and assumes flat diffuse priors for the fixed effects and minimally informative inverse Wishart priors (using MQL1 random parameter estimates) for the random parameters at level 2 and inverse Gamma priors at level 1 (Appendix 2.5). The iterated bootstrap uses 500 replicates per bootstrap set and 10 sets.

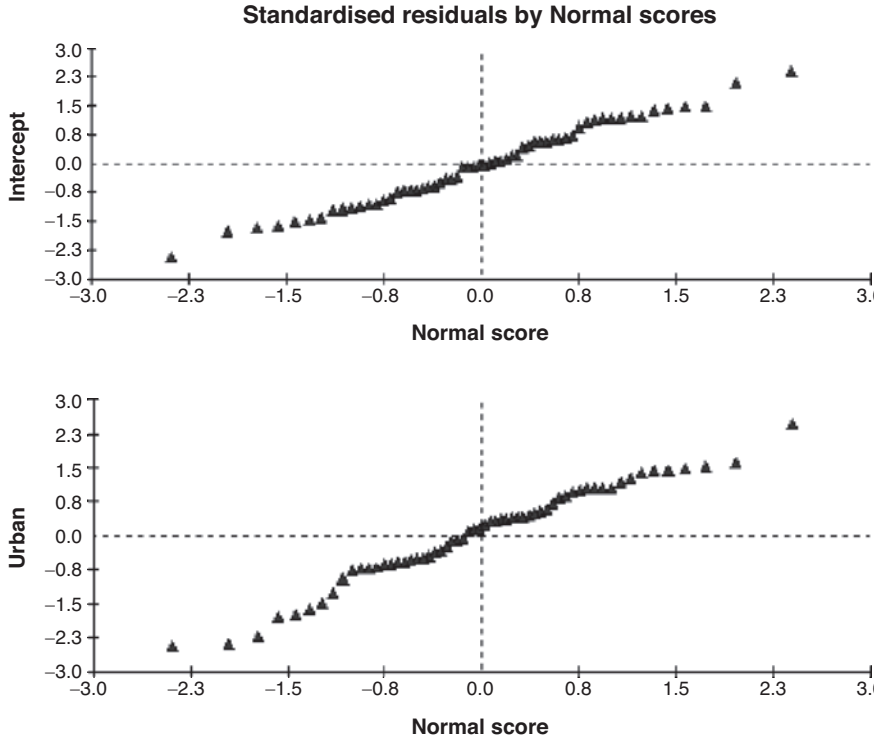


Figure 4.1 Residual plots for analysis B in Table 4.3.

The maximum likelihood deviance criterion for the urban random effects is judged using a chi-bar test and is highly significant and is rather larger than the Wald test for PQL2 estimates. For MCMC the DIC change (Chapter 2) is also substantial, indicating an improvement from adding urban random effects. The parameter estimates for all three methods are similar, and in particular, the variance estimates are somewhat larger than those for PQL2 and substantially larger than for MQL1.

4.3.2 Modelling school segregation

In Chapter 3, we discussed the modelling of the variance structure at different levels. We now give an example where the principal interest of the analysis lies in relating the variation of the response to explanatory variables.

A concern of educators is the extent to which certain minority groups, defined by such factors as ethnicity or social background, may become over-concentrated in some schools. There is a long history of attempts to measure what has come to be termed 'segregation'. A useful summary is given by Allen and Vignoles (2007) who discuss different ways of characterising the variability across institutions. In

Table 4.4 Estimates for model of Table 4.3 using alternative estimation procedures. Standard errors in brackets (see text).

Fixed parameter	Maximum likelihood (simulated)	MCMC	Iterated bootstrap
Intercept	−1.72	−1.72	−1.71
Age*	−0.026 (0.008)	−0.027 (0.008)	−0.027 (0.009)
Urban	0.820 (0.170)	0.805 (0.189)	0.813 (0.182)
1 child	1.138 (0.160)	1.157 (0.157)	1.122 (0.153)
2 children	1.360 (0.177)	1.376 (0.174)	1.356 (0.169)
3+ children	1.358 (0.183)	1.384 (0.180)	1.354 (0.178)
Random parameter			
σ^2_{u0} (intercept)	0.390 (0.120)	0.418 (0.137)	0.403 (0.113)
σ_{u01}	−0.402 (0.182)	−0.432 (0.176)	−0.426 (0.164)
σ^2_{u1} (urban)	0.654 (0.362)	0.738 (0.303)	0.714 (0.300)
Deviance criterion for addition of urban random effect. (DIC for MCMC)	15.8	22.6	

*Age is centred at the mean age of 29.56. The base categories for other variables are 'rural' and 0 children. RIGLS estimation.

fact, all of these different 'indices' can be derived from the parameters of a particular multilevel model, fitted to the data. We shall consider the case of social disadvantage as measured (rather crudely) by the proportion of students in a school who are eligible to receive free school meals (FSM) at lunchtime. For reasons to do with catchment area or selection policies, in general schools will differ in the actual proportions eligible for FSM. Interest lies in whether this differs across different regions of a country or changes across time, etc.

We note, to begin with, that even where there is no mechanism responsible for schools differing in these proportions, random fluctuations will lead to an *observed* difference in proportions. Thus, our aim is to set up a statistical model that allows us to study *underlying* differences and a simple such model is given by

$$\begin{aligned} &\text{logit}(\pi_{ij}) = (X\beta)_{ij} + v_j + u_{ij} \\ &p_{ij} \sim bin(n_{ij}, \pi_{ij}), \quad v_j \sim N(0, \sigma^2_v), \quad u_{ij} \sim N(0, \sigma^2_u) \end{aligned} \tag{4.6}$$

where j indexes Local Education Authority (LEA) (School Board), i indexes schools and p_{ij} is the observed proportion in the ij-th school. Our interest centres on the between-school and between LEA variation, as summarised in the variance parameters σ^2_u, σ^2_v.

Table 4.5 Variance estimates (standard errors) for each of five years for English Secondary schools.

Year	Between-school	Between LEA	Total
1994	0.625 (0.016)	0.491 (0.066)	1.116
1995	0.636 (0.016)	0.522 (0.072)	1.158
1996	0.650 (0.016)	0.503 (0.064)	1.153
1997	0.660 (0.017)	0.498 (0.069)	1.158
1998	0.685 (0.017)	0.506 (0.068)	1.191
1999	0.691 (0.017)	0.506 (0.068)	1.197

A full discussion of this model and the results of the following data analysis is given by Goldstein and Noden (2003), who show that (4.6) is indeed a good fit to the observed data and hence a suitable basis for studying segregation. Their main results are summarised in Table 4.5.

We can see that there is an 11 % increase in the variability between-schools (within LEA) over the five-year period with no discernible trend for LEAs. Goldstein and Noden also find that the increase in segregation is greater for those LEAs that ran a selective secondary school system where students' secondary school was determined, at least partly, on the basis of their prior educational achievement; the authors offer various interpretations of these findings. Leckie *et al*. (2010) show how the traditional segregation indices can be derived more efficiently as functions of estimated model parameters.

4.4 Models for multiple response categories

In this section, we extend the model for a single proportion as outcome to the case of a set of proportions; for example, the proportions voting for three different political parties. The response is now multivariate and we can define a generalisation of the ordinary logit model to define a multivariate logit as follows for a simple 2-level variance components model

$$\log\left(\frac{\pi_{ij}^{(s)}}{\pi_{ij}^{(t)}}\right) = \beta_0^{(s)} + \beta_1^{(s)} x_{ij} + u_j^{(s)}, \quad s = 1, \ldots, t-1 \tag{4.7}$$

where there are t response categories. Choosing one category (say t) as the base category avoids redundancy and a singular covariance matrix and hence the need to introduce generalised inverses into the estimation. There are cases, however, where this procedure is inappropriate and we discuss these below. Thus, (4.7) specifies the model for each of the remaining $t-1$ categories with $\sum_{h=1}^{t} \pi_{ij}^{(h)} = 1$. When $t = 2$ this

reduces to the ordinary logit model. We can also define a multivariate complementary log-log link as shown in Table 4.1, which again reduces to the ordinary log-log model when $t = 2$ (see Chapter 7 for more detail on the corresponding probit link).

We treat the $t - 1$ response categories as a multivariate response vector (see Chapter 6 for details about how multivariate response models are specified). In this case we use dummy variables with no variation at level 1 and the true level 1 covariance matrix specified at level 2. Consider, for example, a single level model where individuals have three response categories, $t = 3$. We specify a bivariate model where level 2 describes the between-individual variation. If we make the standard assumption that the observed response proportions follow a multinomial distribution then the level 2 covariance matrix has the form

$$n_{ij}^{-1}\begin{pmatrix} \pi_{ij}^{(1)}(1-\pi_{ij}^{(1)}) & & & \\ -\pi_{ij}^{(1)}\pi_{ij}^{(2)} & . & & \\ . & . & & \\ . & & . & \\ -\pi_{ij}^{(1)}\pi_{ij}^{(t-1)} & . \quad . \quad . & & \pi_{ij}^{(t-1)}(1-\pi_{ij}^{(t-1)}) \end{pmatrix} \tag{4.8}$$

where n_{ij} is the total number of responses over all categories.

We can create the covariance structure (4.8) as follows. Define the explanatory variables

$$\begin{aligned} z_{1ij} &= \sqrt{\pi_{ij}/n_{ij}}, \quad z_{2ij} = \pi_{ij}/\sqrt{2n_{ij}} \\ z_{3ij} &= -\pi_{ij}/\sqrt{2n_{ij}}, \quad \pi_{ij} = \{\pi_{ij}^{(s)}\} \end{aligned} \tag{4.9}$$

and specify Z_1 to have a random coefficient at level 1 with variance constrained to 1.0 and Z_2, Z_3 to have random coefficients at level 2 constraining their variances to zero and their covariance to 1.0. This produces the structure (4.8); extra multinomial variation can be achieved by allowing the variance and covariance to be different from 1.0 but constraining them to be equal. Level 3 will define variation between higher level units, for example, schools.

The response vector itself is not restricted to a single classification. Thus, suppose we are studying individual grades in examinations. If we had two responses each with three categories, this produces nine response categories of which just one contains the value 1 for each individual. A 'main effects' model extension to (4.7) would express the probability of any particular combination of first and second preferences as an additive function of a term for the first and for the second exam grade as follows

$$\log\left(\frac{\pi_{ij}^{(s=s_1,s_2)}}{\pi_{ij}^{(t)}}\right) = \beta_0^{(s_1)} + \beta_0^{(s_2)} + \beta_1^{(s_1)}x_{1ij} + \beta_1^{(s_2)}x_{2ij} + u_j^{(s_1)} + u_j^{(s_2)},$$
$$s = 1, \ldots, \mathrm{t} - 1$$

For the random parameters it might be reasonable to attempt to fit a model where the covariances between the $u_j^{(s_1)}$, $u_j^{(s_2)}$ were zero in order to reduce the number of random parameters in the model.

To see how we can interpret the parameters of these models, we write, from (4.7)

$$\log\left(\pi_{ij}^{(r)}/\pi_{ij}^{(s)}\right) = \left(\beta_0^{(r)} - \beta_0^{(s)}\right) + \left(\beta_1^{(r)} - \beta_1^{(s)}\right)x_{ij} + \left(u_j^{(r)} - u_j^{(s)}\right) \tag{4.10}$$

so that a unit change in x_{ij} multiplies the ratio of the r-th and s-th response probabilities by $\exp(\beta_r^{(r)} - \beta_1^{(s)})$. Likewise a difference of d in the residuals or in the intercept terms multiplies this ratio by e^d.

This formulation of the multicategory response model is adequate for models such as (4.7) where coefficients are fitted for each response category (except the base). There are other models, however, where we may wish to fit a function defined *across the categories*. This will often be the case when there are a large number of ordered categories where we wish to study linear, quadratic, etc., trends across the categories, although, as we point out in Section 4.2, there will often be more satisfactory procedures for such cases based upon consideration of the *cumulative* probabilities $\pi_{ij}^{(1)}, \pi_{ij}^{(1)} + \pi_{ij}^{(2)}, \ldots$.

Where we do wish to treat the categories symmetrically and define a function across the response categories, we replace the intercept term $\beta_0^{(s)}$ in (4.7) by such a function. If we assume a linear function then (4.7) can be written as

$$\log\left(\frac{\pi_{ij}^{(s)}}{\pi_{ij}^{(t)}}\right) = \gamma_0 + \gamma_1 w^{(s)} + (\beta_0 + \beta_1 w^{(s)})x_{ij} + u_j^{(s)}, \quad s = 1, \ldots, t-1 \tag{4.11}$$

where $w^{(s)}$ is the score assigned to category s. We might also wish to structure the level 2 variation, for example, writing $u_j^{(s)} = u_{0j} + u_{1j}w^{(s)}$. Such a model will be especially useful when the number of categories becomes large.

In (4.11), the choice of base category is no longer irrelevant since the score assigned to this category does not appear in the model. We can avoid this difficulty by defining the multivariate logistic over all the response categories ($s = 1, \ldots, t$); in (4.11), the level 2 resulting covariance matrix will not be singular so long as the set of response category probabilities is predicted using fewer responses than there are categories. An alternative formulation, using the Poisson with a log-link function as described below, will often be more convenient.

As with the case of a single proportion as outcome, we have a choice of quasi-likelihood, maximum likelihood or MCMC estimation procedures. With multiple responses, however, the biases associated with MQL can be very large for (0,1) responses and can also be affected by choice of base category.

4.5 Models for counts

Instead of using a set of proportions as the response we can consider the underlying event counts as the set of responses. Thus, for example, in the single exam grade case, suppose we classify individuals by three initial achievement categories and by gender. In each of the six cells within each level 2 unit, which is now, say, a school, we have counts of the numbers in each of the three grades, which yields 18 counts. The expected number of individuals in each grade can be written

$$m_{sij} = M_j \pi_{ij}^{(s)}$$

where s indexes the grades, i indexes the six cells within each level 2 unit and M_j is the number of individuals in the j-th school. Our inferences are therefore conditional on these totals. We write, corresponding to (4.7)

$$\log(m_{sij}) = \log(M_j) + \beta_0^{(s)} + \beta_1^{(s)} x_{ij} + u_j^{(s)}, \quad s = 1, \ldots t \tag{4.12}$$

The term $\log(M_j)$ is a fixed part offset and when using such offsets it may be better to centre them about their mean, in order to avoid numerical instabilities. Corresponding to the multinomial assumption, we now make the assumption of a Poisson distribution for the *observed* counts n_{sij}, which are assumed conditionally independent with

$$E(n_{sij}) = m_{sij}, \quad \text{var}(n_{sij}|m_{sij}) = m_{sij}$$

We now have a 2-level model where at level 2 we have the school and the level 1 units are the set of counts for the classification of grade by initial achievement category and gender. A basic additive model will have explanatory variables, consisting of an intercept, two dummy variables for grade, two dummy variables for initial achievement and one for gender. We might also wish to include interactions between grade and initial achievement and grade and gender.

The level 1 variation is specified using the predicted number for each level 1 unit and the estimation follows the same pattern as for the binomial model, using the corresponding expressions given in Appendix 4.1. The level 1 random part will be defined by a dummy variable equal to the square root of the predicted count and with variance constrained to one where a Poisson distribution is assumed.

There are some applications where the response is a count and we do not require an offset, or where the offset is effectively constant. For example, if we were interested in the number of times individuals visited their general practitioner or physician in a year, we could collect data over a one-year period for all individuals and study the variation in counts across practitioners (level 2) according to individual, practitioner and area characteristics.

There are variations on the Poisson distribution assumption which we may wish to use. For example, the negative binomial distribution can be obtained from a process whereby the response is generated by counting the number of incidents for each level 1 unit; and where, conditional on the fitted explanatory variables and higher level

terms, the mean count for each level 1 unit has a Gamma distribution with index v. This leads us to consider level 1 variance functions of the general form $k_1 m + k_2 m^2$, where $k_1 = 1$ gives the negative binomial distribution with $k_2 = 1/v$. We could add further terms or consider even a nonlinear function.

As with binomial distribution models we can use maximum likelihood (Appendix 4.2) or MCMC estimation (Appendix 4.3).

4.6 Ordered responses

In the exam grade example of the previous section, we ignored the fact that the scale was ordered. Such response scales are common and are sometimes analysed by assigning scores, then treating them as if they were continuous. While this may often be satisfactory, there are situations – for example, where the distribution is very skew – where such a procedure is questionable. One possible alternative, mentioned in Section 4.5, is to assign scores to the categories of the response variable and then carry out an analysis based upon the multinomial or Poisson model, using all the response categories in the analysis. Such a procedure, like the continuous response model, typically relies on choosing a suitable scoring system. Another option is to assign scores by minimising a measure of between-unit disagreement, as in correspondence analysis or dual scaling (Greenacre, 1984; Goldstein, 1987 c). We will now look at procedures which avoid the arbitrariness of assumptions involved when assigning numerical scores.

To exploit the ordering, we can base our models upon the *cumulative* response probabilities rather than the response probabilities for each category. We define these as

$$E(y_{ij}^{(s)}) = \gamma_{ij}^{(s)} = \sum_{h=1}^{s} \pi_{ij}^{(h)}, \quad s = 1, \ldots, t-1 \tag{4.13}$$

where $y_{ij}^{(s)}$ are the observed proportions out of a total n_{ij} and s now indexes the ordered cumulative categories. If we assume an underlying multinomial distribution for the category probabilities the cumulative proportions have a covariance matrix given by

$$\operatorname{cov}\left(y_{ij}^{(s)}, y_{ij}^{(r)}\right) = \gamma_{ij}^{(s)}\left(1 - \gamma_{ij}^{(r)}\right)/n_{ij}, \quad s \leq r \tag{4.14}$$

We can therefore fit models to these cumulative proportions (or counts, conditional on a fixed total) in the same way as with the multinomial response vector, substituting the covariance matrix (4.14) for (4.8). A discussion of these and related models is given in McCullagh and Nelder (1989). A common model choice is the *proportional odds* model, which uses a logit link, namely

$$\gamma_{ij}^{(s)} = \{1 + \exp -[\alpha^{(s)} + (X\beta)_{ij}]\}^{-1} \tag{4.15}$$

which implies that increasing values of the linear component are associated with increasing probabilities with increasing s. We also require $\alpha^{(1)} \leq \alpha^{(2)} \ldots . \leq \alpha^{(t-1)}$.

Another choice is the *proportional hazards* model, which uses a log-log link to give

$$\gamma_{ij}^{(s)} = \{1 - \exp[-\exp(\alpha^{(s)} + (X\beta)_{ij})]\} \tag{4.16}$$

An important special case of these models is where the categories are ordered in time so that $\alpha^{(s)}$ can be modelled as a function of time, and satisfying the above order relationship among these parameters. Some choices would be

$$\alpha^{(s)} = \delta \log(t_s), \quad \alpha^{(s)} = \delta t_s \tag{4.17}$$

Such a model might be used in developmental studies where individuals pass through a set of time-ordered stages. In studies of children, for example, it is possible to identify 'milestones' of development, starting with none; all have been passed when developmental 'maturity' is reached. A repeated measures study would count the number passed at each time point, so yielding a cumulative proportion in relation to time and other covariates. We would then be able to fit a 2-level model with variation between individuals involving any of the parameters in (4.15), (4.16) or (4.17). We return to such models when we consider survival or event history models in Chapter 11.

Another example of longitudinal discrete response data is where, at each measurement occasion, we have a vector of ordered categorical responses and each individual in the study responds to one category. The cumulative response vector for each individual at each occasion then contains zero for each response category less than the category to which the individual responds and a one for that category and each higher one. We can model the time dependence within the set of explanatory variables X, and would normally wish to include the possibility of interactions between the $\alpha^{(s)}$ and time. In such a model, the basic covariance structure given by (4.14) represents the between-occasion covariation. Thus, although the data structure is represented by level 1 as the categories, level 2 as occasion and level 3 as individual, the higher level variation is only estimated at level 3. This can be compared to the simple binary response model where the binomial response variance is that between occasions, and the structure defines occasion as level 1 and individual as level 2 since there is a single response for each occasion. We also note that similar considerations apply to all the multicategory response models, with higher level variation estimated at level 3 and above.

4.7 Mixed discrete-continuous response models

An extension of the multivariate models to be considered in Chapter 6 is where some of the responses are continuous and some are discrete. In a repeated measures study, we may have a response which is the (discrete) maturity stage that an individual

has reached, as well as continuous measurements such as height and weight; we may wish to treat, say, the maturity stage as the response, conditional on height and weight and further covariates, including age. In other situations, for example, if we are interested in prediction systems, then we would wish to consider all the measurements as responses, conditional on covariates. In another example, suppose we have measurements on smoking habit, including whether someone smoked and if so at what rate. We can consider this as a bivariate response model where each individual has a binary response for whether or not they smoke and if they do a further response for the number smoked per day.

We develop the model for the case of individual smoking habits with one binary and one continuous response, and then look at the more general case of several binary responses. The extension to several responses of each type is straightforward, as is the extension to multicategory responses and count data.

As in the standard multivariate multilevel model, we have no variation at level 1; at level 2, that of the individual, indexed by *i,* we have a binomial variance associated with the smoking/no smoking response and a between-individual variance for the number smoked. The variance for the binary response is the usual binomial variance and that for the continuous response is a parameter to be estimated. At the higher level, indexed by *j*, the variances and covariances will be defined in the standard fashion. For a 2-level model with individuals nested within, say, households we write the model in two parts. For the binary response probability

$$\begin{aligned} \text{logit}(\pi_{ij}) &= (X_1\beta_1)_{ij} + u_{1j} \\ y_{ij} &\sim \textit{binomial}(1, \pi_{ij}) \end{aligned}$$

and for the continuous response

$$y_{ij} = (X_2\beta_2)_{ij} + u_{2j} + e_{ij}$$

with

$$e_{ij} \sim N(0, \sigma_e^2), \quad \begin{pmatrix} u_{1j} \\ u_{2j} \end{pmatrix} \sim N(0, \Omega_u), \quad \Omega_u = \begin{pmatrix} \sigma_{u1}^2 & \\ \sigma_{u12} & \sigma_{u2}^2 \end{pmatrix} \tag{4.18}$$

In the general case, an individual can have any combination of responses – as in the maturity example – and the individual level covariance will have the form of an (adjusted) biserial covariance $(1 - \hat{\pi}_{jk})\hat{\pi}_{jk}(\hat{y}_{1jk} - \hat{y}_{2jk})$, where $\hat{\pi}_{jk}$ is the estimated probability of a positive response and $\hat{y}_{1jk}$, $\hat{y}_{2jk}$ are respectively the predicted values of the continuous response for a positive and negative binary response. We can fit this using an extra covariance term in the model at the individual level, constrained to have the above value. If we assume that $\hat{y}_{1jk} - \hat{y}_{2jk}$ is constant, then we can fit this term by defining a further explanatory variable equal to the existing variable, defining the binomial variation at the individual level, and fitting just a covariance term between this further explanatory variable and the existing binomial explanatory variable. This gives the required estimate of $\hat{y}_{1jk} - \hat{y}_{2jk}$. In the smoking

example, since non-smokers do not have any number smoked, this covariance term does not exist.

In the next section, we look at a more flexible alternative formulation that allows us to model categorical variables in terms of underlying continuous ones.

4.8 A latent normal model for binary responses

In many cases, a binary response can be thought of as being derived by choosing a threshold from an underlying continuous distribution so that the response is 1 if above the threshold and 0 if below. This might arise in assigning pass/fail grades for an examination based on a mark cut-off. Likewise, a series of ordered grades may be derived from an underlying mark scale. Another example is in the categorisation of illness severity, where there is no actual underlying continuous scale but interest may lie in constructing one from observed categories. A similar situation might arise with attitude measurement.

Suppose that we have a variance components 2-level model for the underlying continuous variable written as

$$y_{ij} = (X\beta)_{ij} + u_j + e_{ij} \tag{4.19}$$

and suppose a positive value occurs when $y_{ij} > 0$. We then have

$$\Pr(y_{ij} > 0) = pr(e_{ij} > -[(X\beta)_{ij} + u_j]) \tag{4.20}$$

Now if we assume $e_{ij} \sim N(0, 1)$, (equivalent to fixing the lowest level variance to be binomial) the probability in (4.20) is

$$\int_{-[(X\beta)_{ij}+u_j]}^{\infty} \phi(t)dt = \int_{-\infty}^{[(X\beta)_{ij}+u_j]} \phi(t)dt$$

where $\phi(t)$ is the probability density function (pdf) of $N(0,1)$, which is in fact just the probit link function. In Appendix 4.3, we show how we can estimate the parameters of this model, where the 'latent' random variable has a normal distribution, using MCMC, sampling from the underlying level 1 N(0,1) distribution.

We have an analogous model for the logit link function. The logistic distribution has the following density function

$$f(x) = \frac{\exp(x)}{[1 + \exp(x)]^2} \tag{4.21}$$

which has zero mean and variance 3.290 and where the required cumulative function is obtained as

$$\int_{-Y}^{\infty} f(x)dx = [1 + \exp(-Y)]^{-1} \tag{4.22}$$

which is just the usual logit link function.

We can also model an underlying continuous distribution for the log-log link. We have a Gumbel distribution with density function

$$f(x) = \exp(-x)(\exp -[\exp(-x)]) \tag{4.23}$$

which has mean -0.577 and variance 1.645. We have

$$\int_{-Y}^{\infty} f(x)dx = 1 - \exp(-\exp(Y)) \tag{4.24}$$

which is just the complementary log-log link function.

We can carry out analogous estimation for ordered data, unordered data and count data, as detailed in Chapter 7, where we also discuss multivariate models with mixtures of normal and discrete responses; these models use the probit link function that leads to underlying multivariate normal distributions.

4.9 Partitioning variation in discrete response models

In Chapter 3, we pointed out that in random coefficient models the variance partition coefficient (VPC) was a function of the predictor variables with random coefficients. In discrete response models, the computation and interpretation of the VPC is more complex. We now discuss a model with a binary response, but our remarks apply more generally to models for proportions, for different non-identity link functions and also where the response is a count, in fact, to any nonlinear model. For a (0,1) response, Model (4.1) can be written

$$\begin{aligned} &E(y_{ij}) = \pi_{ij} = f(\beta_0 x_0 + \beta_1 x_{1ij} + u_{0j}) \\ &y_{ij} \sim \textit{Bernoulli}(\pi_{ij}) \\ &u_{0j} \sim N(0, \sigma^2_{u0}) \end{aligned} \tag{4.25}$$

Unlike in the normal case, the level 1 variance depends on the expected value, $\text{var}(y_{ij}) = \pi_{ij}(1 - \pi_{ij})$ and the fixed predictor in the model depends on the value of x_1. Therefore, as we are considering a function of the predictor variable x_1, a simple VPC is not available, even though there is only a single level 2 variance. Furthermore, the level 2 variance, σ^2_{u0}, is measured on the logistic scale so is not directly comparable to this level 1 variance.

If we do wish to produce a measure that is the counterpart of that for continuous responses, then the following procedures can be considered.

4.9.1 Model linearisation (Method A)

Using a first order Taylor expansion, as described in Appendix 4.1, we can write (4.25) in the form

$$y_{ij} = (\beta_0 + \beta_1 x_{1ij}) f'_{ij} + u_j f'_{ij} + e_{ij}\sqrt{\pi_{ij}(1-\pi_{ij})}$$
$$\text{var}(e_{0ij}) = 1$$

where we evaluate π_{ij} at the mean of the distribution of the level 2 random effect, that is, for the logistic model

$$\pi_{ij} = \exp(\beta_0 + \beta_1 x_{1ij})[1 + \exp(\beta_0 + \beta_1 x_{1ij})]^{-1}$$
$$f'_{ij} = \pi_{ij}[1 + \exp(\beta_0 + \beta_1 x_{1ij})]^{-1} \tag{4.26}$$

so that, for a given value of x_1 we have

$$\text{var}(y_{ij}|x_{1ij}) = \sigma^2_{u0}\pi^2_{ij}[1 + \exp(\beta_0 + \beta_1 x_{1ij})]^{-2} + \pi_{ij}(1-\pi_{ij})$$

and

$$\tau = \sigma^2_{u0}\pi^2_{ij}[1 + \exp(\beta_0 + \beta_1 x_{1ij})]^{-2}$$
$$\times\{\sigma^2_{u0}\pi^2_{ij}[1 + \exp(\beta_0 + \beta_1 x_{1ij})]^{-2} + \pi_{ij}(1-\pi_{ij})\}^{-1}$$

where sample estimates are substituted.

4.9.2 Simulation (Method B)

This method is general and can be applied to any non-linear model without the need to evaluate an approximating formula. It is computationally reasonably fast and can be made to yield as accurate a result as desired by increasing the number of simulations. It consists of the following steps:

1. From the fitted model, say (4.25), simulate a large number m (say 5000) values for the level 2 residual from the distribution $N(0, \sigma^2_{u0})$, using the sample estimate of the variance.

2. For a particular chosen value(s) of x_1 compute the m corresponding values of π_{ij} (π^*_{ij}) using (4.25). For each of these values compute the level 1 variance

$v_{1ij} = \pi^*_{ij}(1 - \pi^*_{ij})$. In the case, say, of a Poisson response with a log link we would compute the level 1 variance as π^*_{ij}.

3. The coefficient is now estimated as

$$\tau = v_2(v_2 + v_1)^{-1}$$
$$v_2 = \text{var}(\pi^*_{ij}), \quad v_1 = E(v_{1ij})$$

4.9.3 A binary linear model (Method C)

As a very approximate indication for the VPC we can consider treating the (0,1) response as if it were a normally distributed variable and estimate the VPC as in that case. This will generally be acceptable when the probabilities involved are not extreme, but if any of the underlying probabilities are close to 0 or 1, this model would not be expected to fit well, and may predict probabilities outside the (0,1) range.

4.9.4 A latent variable approach (Method D)

As in Section 4.8, we may consider the observed (0,1) responses as arising from an underlying continuous variable so that a 1 is observed when a certain threshold is exceeded, otherwise a 0 is observed. For the logit model we have the underlying logistic distribution given by (4.21) and (4.22)

Since the variance for the standard logistic distribution is $\pi^2/3 = 3.29$ we take this to be the level 1 variance and both the level 1 and level 2 variances are on a continuous scale. We now simply calculate the ratio of the level 2 variance to the sum of the level 1 and level 2 variances to obtain the VPC. A similar computation can be used for the probit and log-log link functions. In Chapter 7, we describe a latent normal model for the Poisson distribution so that a similar procedure can be used where the level one variance is fixed at 1.0.

This approach may be reasonable where, say, a (0,1) response is derived from a division of an underlying continuum such as a pass/fail response based upon a continuous mark scale, but would seem to have less justification when the response is truly discrete, such as mortality or voting. (See Snijders and Bosker, 1999, Chapter 14, for a further discussion.)

If we have fitted multiplicative extra-binomial variation (Section 4.1), then in the above approaches the level 1 variance is multiplied by the estimated scaling parameter. For additive extra-binomial variation this variance is added to the level 1 variance.

For the case of an ordered response with p categories, corresponding to methods A–C, we can treat each cumulative proportion in the same way as binary data, yielding $p - 1$ VPC estimates. For D we can use the latent normal approach, as described in Chapter 7, which corresponds to that for a binary response.

4.9.5 An example of VPC calculations

We illustrate the procedures using some data on voting patterns (Heath *et al*., 1996). The model response is whether or not, in a sample of 800 respondents, they expressed a preference for voting conservative ($y_{ij} = 1$) in the 1983 British general election. Several covariates were originally fitted but for simplicity we fit only an intercept model, namely

$$
\begin{aligned}
&E(y_{ij}) = \pi_{ij} = \exp(\beta_0 + u_{0j})[1 + \exp(\beta_0 + u_{0j})]^{-1} \\
&y_{ij} \sim \textit{Bernoulli}(\pi_{ij}) \\
&u_{0j} \sim N(0, \sigma^2_{u0})
\end{aligned}
$$

The fitted model parameters are as follows (using PQL2 with IGLS):

$$\hat{\beta}_0 = -0.256, \quad \hat{\sigma}^2_{u0} = 0.142$$

The VPC estimates are given in Table 4.6.

There is, as expected, good agreement for methods A, B and C with that for D somewhat larger. Now, however, we choose a more extreme case. In Chapter 9 of Rasbash *et al*. (2008), these data are fitted with four predictor variables measuring political attitudes. The scales of these predictor variables are constructed such that lower values correspond to left wing attitudes. If we take the set of values corresponding to the fifth percentile point of each of the four predictors, then the predicted value on the logit scale is approximately –2.5, corresponding to a low probability of voting Conservative of 0.076. At this value of the predictor, the respective VPCs for methods A and B are 0.0096 and 0.0111, which are again in reasonable agreement. For method C, fitting the four predictor variables we obtain a value of 0.0257, which is very different. For method D, the residual level 2 variance does not change very much from the previous model, and we obtain a VPC of 0.047, which is also different to the others. We note that the VPC for methods C and D do not depend on the value of the linear predictor.

If one wishes to make inferences on an underlying continuous scale, then Method D is appropriate. The choice of whether to report on the probability scale or an

Table 4.6 Level 2 and level 1 estimated variances with variance partition coefficient (VPC).

	Method			
	A	B	C	D
Level 2 variance	0.0086	0.0083	0.0088	0.142
Level 1 variance	0.246	0.238	0.237	3.290
VPC	0.034	0.034	0.036	0.043

underlying continuous scale will depend on the application; in the present example, it may seem more natural to report directly on the probability scale rather than on an assumed underlying continuous scale of 'propensity' to vote Conservative. We can obtain an interval estimate for any of our estimates via the bootstrap or using the results from an MCMC estimation run. Goldstein *et al*. (2002) give more details of these procedures, including MLwiN macros for computing them.

Appendix 4.1 Multilevel generalised linear model estimation

4.1.1 Approximate quasilikelihood estimates

We consider a single nonlinear model of the form

$$y_{ij} = f(X_{2ij}\beta_2 + Z_{2ij}u_{2j}) + Z_{1ij}C_{2ij} \tag{4.1.1}$$

At the $(t+1)$-th iteration we expand the nonlinear function f in (4.1.1) for both fixed and random parts as follows using a Taylor series

$$f_{ij}(H_t) + X_{2ij}(\beta_{2,t+1} - \beta_{2,t})f'_{ij}(H_t) + (Z_{2ij}u_{2j})f'_{ij}(H_t) + (Z_{2ij}u_{2j})^2 f''_{ij}(H_t)/2 \tag{4.1.2}$$

in terms of parameter values estimated at the t-th iteration. The first two terms of (4.1.2) update the fixed part of the model and in the special case of a single level model provides the usual updating function in maximum likelihood estimation. The quantity $f_{ij}(H_t) - X_{2ij}\beta_{2,t}f'_{ij}(H_t)$ is treated as an offset to be subtracted from the response variable. The third term defines a linear random component based on the explanatory variables transformed by multiplying by the first differential. We need to specify H_t and then consider the distribution of this term. The level 1 term in (4.1.1) is fitted in the usual way; the choice of Z_{1ij} depending on the nature of the response, as discussed in Section 4.1.

If we choose $H_t = X_{2ij}\beta_{2,t}$, this is equivalent to carrying out the Taylor expansion around the fixed part predicted value. If we choose $H_t = X_{2ij}\beta_{2,t} + Z_{2ij}\hat{u}_{2j}$, this expands around the current predicted value for the ij-th unit and we replace the last two terms of (4.1.2) by

$$Z_{2ij}(u_{2j} - \hat{u}_{2j})f'_{ij}(H_t) + Z_{2ij}(u_{2j} - \hat{u}_{2j})^2 f''_{ij}(H_t)/2$$

We thus have the further offset from the linear term to be added to the response

$$(Z_{2ij}\hat{u}_{2j})f'_{ij}(H_t)$$

A discussion of these approaches in the context of multilevel generalised linear models is given by Breslow and Clayton (1993). Wolfinger (1993) synthesises some of the literature based upon this 'predictive' approach. All these methods use only the first order terms in (4.1.2).

From the second line of (4.1.2), where the Taylor expansion is about zero, we have

$$E(Z_{2ij}u_{2j}) = 0, \quad E(Z_{2ij}u_{2j})^2 = \sigma^2_{zu}, \quad \sigma^2_{zu} = Z_{2ij}\Omega_u Z^T_{2ij} \tag{4.1.3}$$

To incorporate the second order terms we treat $\sigma^2_{zu} f''(H_t)/2$ as an additional offset in the fixed part and in the random part of the model and we need to consider the variation of the last term of (4.1.2). If we assume normality then all third moments, formed from the product of the last two terms of (4.1.2), are zero and we have

$$\text{var}(Z_{2ij} u_{2j})^2 = 2\sigma^4_{zu} \tag{4.1.4}$$

so that we need to define the additional random variable $Z^*_u = \sigma^2_{zu} f''(H_t)/\sqrt{2}$ with variance constrained to be equal to 1.0. Equivalently we can form $Z^*_u Z^{*^T}_u$ as an offset for the response vector $vec(\tilde{Y}\tilde{Y}^T)$ in the estimation of the random parameters.

Having modified the response variable by removing the necessary offsets we are left in the fixed part with a modified response, say Y' with a modified explanatory variable matrix, say X'. We do likewise for the random part of the model and then carry out a standard iterative procedure, updating the differential functions at each iteration.

Where the Taylor expansion is taken about the current values of the residuals, we require $E(Z_{2ij}(u_{2j} - \hat{u}_{2j}))^2$ which leads to the 'conditional' or 'comparative' variances of the residuals as described in Appendix 2.2, so that we substitute these in the above expressions for the fixed and random offsets.

To estimate residuals we note that, having adjusted the response using the offsets, we have on the right hand side of the model, for the Taylor expansion about zero, the fixed part together with the random terms

$$(Z_{2ij} u_{2j}) f'_{ij}(H_t) + ((Z_{2ij} u_{2j})^2 - \sigma^2_{zu}) f''_{ij}(H_t)/2$$

Each residual and its square appear in this expression, and since third order moments are zero, we can apply the usual linear estimation for the residuals as described in Appendix 2.2. The weight matrix V is based upon both the linear and quadratic terms of the above expression. We carry out an analogous procedure for the case where the Taylor expansion is based upon the current residual estimates.

The above can be extended in a straightforward way to more than two levels and to multivariate models. For the first order approximation the procedure outlined here is closely related to that given by Lindstrom and Bates (1990) for 2-level repeated measures data who consider a first order expansion about the unit-specific predicted values. Gumpertz and Pantula (1992) consider a variance components model where the fixed part predictor is nonlinear.

For generalised linear models Waclawiw and Liang (1993) consider a generalised estimating equations (GEE) approach (see Chapter 2), using a unit-specific predictor. A full likelihood based method for a repeated measures model with binary responses is described by Garret *et al*. (1993) and issues arising from applications to repeated measures data are discussed in Chapter 5.

For small samples, as discussed in Appendix 2.1, we should use the (RIGLS, REML) procedure. As discussed in Chapter 4, for certain data configurations this

procedure can produce biased estimates of the fixed and random parameters. In Appendix 4.2, we discuss maximum likelihood estimation and bootstrapping.

4.1.2 Differentials for some discrete response models

The Logit – Binomial model

$$
\begin{aligned}
f &= [1 + \exp(-X\beta)]^{-1} \\
f' &= f[1 + \exp(X\beta)]^{-1} \\
f'' &= f'[1 - \exp(X\beta)][1 + \exp(X\beta)]^{-1}
\end{aligned}
$$

The Logit – Multinomial (Multivariate Logit) model

$$
\begin{aligned}
f^{(s)} &= \exp(X\beta^{(s)}) \left[1 + \sum_{h=1}^{t-1} \exp(X\beta^{(h)}) \right]^{-1}, \quad s = 1, \ldots, t-1 \\
f^{/(s)} &= f^{(s)} \left[1 + \sum_{h=1}^{t-1} \exp(X\beta^{(h)}) \right]^{-1} \left[1 + \sum_{h \neq s} \exp(X\beta^{(h)}) \right] \\
f^{//(s)} &= f^{/(s)}(1 - 2f^{(s)})
\end{aligned}
$$

The Log – Poisson model

$$
\begin{aligned}
f &= \exp(X\beta) \\
f' &= \exp(X\beta) \\
f'' &= \exp(X\beta)
\end{aligned}
$$

The complementary log log – Binomial model

$$
\begin{aligned}
f &= 1 - \exp[-\exp(X\beta)] \\
f' &= (1 - f)\exp(X\beta) \\
f'' &= f'[1 - \exp(X\beta)]
\end{aligned}
$$

Appendix 4.2 Maximum likelihood estimation for multilevel generalised linear models

4.2.1 Simulated maximum likelihood estimation

We have seen that for normally distributed data maximum likelihood (ML) or restricted maximum likelihood (REML) estimates can be obtained readily using the IGLS and RIGLS algorithms. For discrete response or generalised linear models the estimation becomes more complicated. We discuss in Section 4.2.2 how numerical quadrature can be used but that this becomes infeasible for models with large numbers of random coefficients and/or levels. Several simulation based alternatives have been suggested (see, for example, McCulloch, 1997 who gives an example for a logit model). One of these methods uses a stochastic version of the EM algorithm where, instead of obtaining the relevant expectations in the E step, the expectations are estimated by generating draws from the estimated distribution of the random effects using an MH algorithm with a suitable proposal distribution and averaging over the draws. Another method is to use an iterative (scoring) algorithm for a generalised linear model where, for the weights in the estimation of updated parameters, we use expectations calculated again as averages over MH sampled values.

A useful general procedure, that is relatively straightforward to implement, is simulated maximum likelihood where the likelihood is evaluated directly by simulation. We can write the full likelihood and corresponding loglikelihood as

$$\begin{aligned} L(\beta, \Omega, U) &= \prod f(Y|U;\beta) f(U;\Omega) \\ \log[L(\beta, \Omega, U)] &= \sum \{\log[f(Y|U;\beta)] + \log[f(U;\Omega)]\} \end{aligned} \tag{4.2.1}$$

This is an example of an extended likelihood (Appendix 2.3) and leads, for example, to the estimation methods of Appendix 4.1 based upon approximating the nonlinear function $f(Y|U;\beta)$.

An exact likelihood approach is to estimate the fixed and random parameters of our model by treating the actual random effects U in (4.2.1) as nuisance parameters, and to integrate them out and work with the *marginal likelihood* given by

$$L(\beta, \Omega) = \int f(Y|U;\beta) f(U;\Omega) dU \tag{4.2.2}$$

where the first term on the right-hand side is the distribution function for the responses conditional on the random effects, or residuals, U. The second term is the distribution function for the random effects. The first term, given U, depends only on the unknown parameters β and the second only on the unknown parameters Ω. Thus, for example, for a 2-level logistic binomial response model where the random effects are assumed to be multivariate normal we have, since the random effects are independent

across units,

$$L(\beta, \Omega) = \prod_j \int \prod_i \{(\pi_{ij})^{s_{ij}}(1-\pi_{ij})^{n_{ij}-s_{ij}}\}\Phi(u_j; \Omega)du_j$$
$$\pi_{ij} = \{1 + \exp(-X_{ij}\beta_j)\}^{-1}, \quad \beta_j = \beta + u_j \tag{4.2.3}$$

where Φ is the multivariate normal density function for the u_j and n_{ij}, s_{ij} are the numbers of trials and successes.

Now (4.2.2) is simply the expected value of $f(Y|U;\beta)$over the distribution of the random effects U. For any given choice of β, Ω we could therefore, in principle, approximate the likelihood by repeatedly sampling sets of random effects from $f(U;\Omega)$ computing $f(Y|U;\beta)$ and forming

$$\sum_{h=1}^{N} f(Y|U_h;\beta)/N \tag{4.2.4}$$

where N is a suitably large number to achieve an acceptable accuracy. The problem is that we do not know the maximum likelihood values of the random parameters to use in the sampling. Suppose, however, that we have a good approximation to Ω, say $\tilde{\Omega}$, and we write (4.2.2) as

$$L(\beta, \Omega) = \int \frac{f(Y|U;\beta)f(U;\Omega)f(U;\tilde{\Omega})}{f(U;\tilde{\Omega})}dU \tag{4.2.5}$$

We can approximate (4.2.5) for the j-th unit by

$$\sum_{h=1}^{N} \frac{f(Y|U_j;\beta)f_h(U_j;\Omega)}{f_h(U_j;\tilde{\Omega})} \Big/ N \tag{4.2.6}$$

where sampling is now with respect to the distribution of the random effects, the importance distribution, given the value $\tilde{\Omega}$. The term

$$\frac{f(U;\tilde{\Omega})}{f(U;\Omega)}$$

adjusts for the different probability density of the importance distribution.

This is known as *importance sampling* and (4.2.6) will provide the required likelihood to any degree of accuracy for β, Ω, for any choice of $\tilde{\Omega}$, although a poor choice will not be efficient. We can improve efficiency by choosing $f(U;\tilde{\Omega})$ so that $U_j \sim N(\hat{U}_j, \hat{\Omega}_{uj})$, where $\hat{U}_j$ is the set of estimated residuals for unit j with covariance matrix Ω_{uj}. These can be estimated from a preliminary sampling run based upon the initial MQL or PQL starting values (see below).

In practice, it is more convenient to work with the loglikelihood so that, for example, for the binomial logit-normal model the loglikelihood is

$$
\begin{aligned}
L &= \sum_j (\ln\{\sum_{h=1}^{N} [\prod_i \{(\pi_{ij}^{(h)})^{s_{ij}}(1-\pi_{ij}^{(h)})^{n_{ij}-s_{ij}}\} \frac{\Phi(u_j^{(h)};\Omega)}{\Phi(u_j^{(h)};\tilde{\Omega})}]/N\}) \\
&= \sum_j (\ln\{\sum_{h=1}^{N} \exp[\sum_i [s_{ij}\ln(\pi_{ij}^{(h)}) + (n_{ij}-s_{ij})\ln(1-\pi_{ij}^{(h)})] \\
&\quad + \ln(\frac{\Phi(u_j^{(h)};\Omega)}{\Phi(u_j^{(h)};\tilde{\Omega})})]/N\}) \qquad (4.2.7) \\
&= \sum_j (\ln\{\sum_{h=1}^{N} \exp[\sum_i [s_{ij}\ln(\pi_{ij}^{(h)}) + (n_{ij}-s_{ij})\ln(1-\pi_{ij}^{(h)})] \\
&\quad + \frac{1}{2} u_j^{(h)^T}(\tilde{\Omega}^{-1} - \Omega^{-1}) u_j^{(h)} + \frac{1}{2}\ln\left(\frac{|\tilde{\Omega}|}{|\Omega|}\right)]\} - \ln(N))
\end{aligned}
$$

where h indexes the sets of random draws and u is a $p \times 1$ vector for each level 2 unit, where p is the number of random coefficients. Note that if $\Omega = 0$, the terms in Ω are omitted. To find the maximum likelihood solution near to $\tilde{\Omega}$, we need to carry out a numerical search over the parameter set β, Ω. Standard routines for maximising general functions can be used and are available in various mathematical and statistical software packages. Some care is required to ensure that the calculations are carried out and stored with sufficient accuracy and double precision arithmetic will usually be necessary. When generating the U the random number seed should be reset to the same value at the start of each set of random draws. This avoids instability due to the stochastic nature of the likelihood estimates. Also, this stochastic element gives a small bias in the estimates of the parameters which decreases as N increases, and is an area for further research. For $\tilde{\Omega}$ we can use second order PQL (PQL2) estimates, or where these may be unobtainable due to convergence problems we can use PQL1 or MQL estimates, possibly together with an iterated bootstrap (Appendix 4.4), to obtain $\tilde{\Omega}$. An approximation to the covariance matrix of the estimated parameters can be obtained from the inverse of the (Hessian) matrix of second derivatives of (minus) the loglikelihood with respect to the parameters at the solution point. This will be available as output from the optimisation routine.

Note that the extension to models with more levels is relatively straightforward. Thus, for a 3-level model, since there is independence across levels, we have for the j-th level 2 unit within the k-th level 3 unit

$$
\begin{aligned}
&\Phi(u_{kj}^{(h)}, u_k^{(h)}; \Omega) = \Phi(u_k^{(h)}; \Omega_3)\Phi(u_{jk}^{(h)}; \Omega_2) \\
&\Omega = \begin{pmatrix} \Omega_3 & \\ & \Omega_2 \end{pmatrix}
\end{aligned}
$$

and for the k-th level 3 unit the third line of (4.2.6) becomes

$$\sum_j (\ln\{\sum_{h=1}^{N} (\exp[\sum_i [s_{ijk}\ln(\pi_{ijk}^{(h)}) + (n_{ijk} - s_{ijk})\ln(1-\pi_{ijk}^{(h)})] + \frac{1}{2} u_{jk}^{(h)^T} (\tilde{\Omega}_2^{-1} - \Omega_2^{-1}) u_{jk}^{(h)} + \frac{1}{2} u_k^{(h)^T} (\tilde{\Omega}_3^{-1} - \Omega_3^{-1}) u_k^{(h)} + \frac{1}{2} \ln\left(\frac{|\tilde{\Omega}_2|}{|\Omega_2|}\frac{|\tilde{\Omega}_3|}{|\Omega_3|}\right)]\} - \ln(N))$$

and these are summed over level 3 units. At the maximum likelihood estimates we also have an estimate of the likelihood itself and this can be used for inference, for example, for comparing two nested models. Likewise, a numerical search procedure can be extended to provide full or profile likelihood confidence intervals (Chapter 2). This approach can be generalised to other distributions, link functions and structures. Thus, for example, for continuous responses we can fit a t-distribution at level 1 or, say, a beta distribution. Likewise various types of link function can be used, including the complementary log-log or probit for binomial data. Multinomial, ordered or unordered data can be fitted using a suitable parameterisation. We can also fit more complex structures such as mixture distributions.

4.2.2 Residuals

The residuals at higher levels are defined as the expected values of the random effects given the data and parameter values. Thus, we have

$$E(U) = \int U f(Y|U;\beta) f(U;\Omega) dU$$

and we obtain the expected value of u as

$$\exp[\sum_j (\ln\{\sum_{h=1}^{N} [\prod_i \{(\pi_{ij}^{(h)})^{s_{ij}} (1-\pi_{ij}^{(h)})^{n_{ij}-s_{ij}}\} u_{j,k}^{(h)} \frac{\Phi(u_j^{(h)};\Omega)}{\Phi(u_j^{(h)};\tilde{\Omega})}]/N\}) - L]$$

and the expected value of the product of two random effects $u_{k_1} u_{k_2}$ as

$$\exp[\sum_j (\ln\{\sum_{h=1}^{N} [\prod_i \{(\pi_{ij}^{(h)})^{s_{ij}} (1-\pi_{ij}^{(h)})^{n_{ij}-s_{ij}}\} u_{j,k_1}^{(h)} u_{j,k_2}^{(h)} \frac{\Phi(u_j^{(h)};\Omega)}{\Phi(u_j^{(h)};\tilde{\Omega})}]/N\}) - L]$$

from which we can obtain the estimated covariance matrix of the residuals. Note that we need to scale the expected values using the likelihood L.

4.2.3 Cross classification and multiple membership models

For a 2-way cross classification we have the log likelihood

$$\ln\{\sum_{h=1}^{N}(\exp[\sum_{j}\sum_{i}[s_{ij}\ln(\pi_{ij}^{(h)})+(n_{ij}-s_{ij})\ln(1-\pi_{ij}^{(h)})]+\frac{1}{2}u_i^{(h)^T}(\tilde{\Omega}_A^{-1}-\Omega_A^{-1})u_i^{(h)}$$
$$+\frac{1}{2}u_j^{(h)^T}(\tilde{\Omega}_B^{-1}-\Omega_B^{-1})u_j^{(h)}+\frac{1}{2}\ln\left(\frac{|\tilde{\Omega}_A|}{|\Omega_A|}\frac{|\tilde{\Omega}_B|}{|\Omega_B|}\right)]\}-\ln(N)$$

where i is associated with classification A and j with classification B. Note that we have to sum over all level 2 cells of the cross classification for each value of h.

For a multiple membership model at level 2 we simulate from a multivariate normal (or t) distribution as above but now for a level 1 unit the level 2 covariance matrix is given by

$$\left(\sum_j w_{ij}^2\right)\Omega_2$$

where the weights w are specific to the level 1 unit. Thus, we form the multiple membership weighted sum $\sum_j w_{ij}u_j = u_{iw}$ say, and use these random effects applied to the above covariance matrix. Again we sum over all level 2 units for each h.

4.2.4 Computing issues

Several computing issues arise. The number of simulation sets N is a matter for further study, although values between 1000 and 2000 have proved adequate on the limited range of examples so far studied. The optimisation procedure is also important and efficient nonlinear search algorithms are necessary. Inequality constraints are needed to ensure positive definite covariance matrices and upper and lower bounds for the fixed parameters may be important to ensure efficiency, as well as good starting values, for example, from a PQL estimation. General search procedures are able to compute search directions numerically but supplying analytical first derivatives will improve speed and reliability. The first derivatives of L with respect to the parameters can be computed using the formulae in Appendix 4.1. Thus, for the logit model, we have the first derivative

$$\frac{\partial L}{\partial\theta}=\frac{1}{\sum_{h=1}^{N}\exp(g_j^{(h)})}\sum_j\frac{\partial}{\partial\theta}\sum_{h=1}^{N}\exp(g_j^{(h)})$$
$$=\frac{1}{\sum_{h=1}^{N}\exp(g_j^{(h)})}\sum_j\sum_h(\exp(g_j^{(h)})\sum_i\{\frac{\partial}{\partial\theta}[s_{ij}\ln(\pi_{ij}^{(h)})+(n_{ij}-s_{ij})\ln(1-\pi_{ij}^{(h)})]\}$$
$$+\frac{\partial}{\partial\theta}[\frac{1}{2}u_j^{(h)^T}(\tilde{\Omega}^{-1}-\Omega^{-1})u_j^{(h)}+\frac{1}{2}\ln\left(\frac{|\tilde{\Omega}|}{|\Omega|}\right)])$$

where $g_j^{(h)} = \sum_i [s_{ij} \ln(\pi_{ij}^{(h)}) + (n_{ij} - s_{ij}) \ln(1 - \pi_{ij}^{(h)})] + \frac{1}{2} u_j^{(h)^T} (\tilde{\Omega}^{-1} - \Omega^{-1}) u_j^{(h)} + \frac{1}{2} \ln\left(\frac{|\tilde{\Omega}|}{|\Omega|}\right)]$

For a fixed coefficient β_k this becomes

$$\frac{1}{\sum_{h=1}^{N} \exp(g_j^{(h)})} \sum_j \sum_h (\exp(g_j^{(h)}) \sum_i \frac{x_{ij,k}}{1 + \exp(X\beta)_{ij}^{(h)}} \times \{[s_{ij}(1 + \exp(X\beta)_{ij}^{(h)}) - n_{ij} \exp(X\beta)_{ij}^{(h)}]\}$$

and for variance and covariance terms we obtain corresponding expressions using

$$\frac{\partial}{\partial\theta}[\frac{1}{2} u_j^{(h)^T} (\tilde{\Omega}^{-1} - \Omega^{-1}) u_j^{(h)} + \frac{1}{2} \ln\left(\frac{|\tilde{\Omega}|}{|\Omega|}\right)] = -\frac{\partial}{\partial\theta}[\frac{1}{2} u_j^{(h)^T} \Omega^{-1} u_j^{(h)} + \frac{1}{2} \ln(|\Omega|)]$$
$$= \frac{1}{2} u_j^{(h)^T} \Omega^{-1} \frac{\partial \Omega}{\partial \sigma_{kl}} \Omega^{-1} u_j^{(h)} - \frac{1}{2} tr(\Omega^{-1} \frac{\partial \Omega}{\partial \sigma_{kl}})$$

where, for example for a bivariate distribution, $\frac{\partial \Omega}{\partial \sigma_1^2} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$, $\frac{\partial \Omega}{\partial \sigma_{12}} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$.

4.2.5 Maximum likelihood estimation via quadrature

We saw above that the likelihood for a 2-level logistic model can be written as the product of terms such as the following for level 2 unit j

$$\int_{-\infty}^{\infty} \prod_i \{(\pi_{ij})^{s_{ij}} (1 - \pi_{ij})^{n_{ij} - s_{ij}}\} f(u_j; \Omega) du_j$$
$$\pi_{ij} = \{1 + \exp(-X_{ij}\beta_j)\}^{-1}, \quad \beta_j = \beta + u_j \tag{4.2.8}$$

where $f(u_j; \Omega)$ is typically assumed to be the multivariate normal density and can be written in the form $\int_{-\infty}^{\infty} P(u_j) f(u_j) du_j$.

Gauss-Hermite quadrature approximates an integral such as the above as

$$\int_{-\infty}^{\infty} P(v) e^{-v^2} dv \approx \sum_{q=1}^{Q} P(x_q) w_q \tag{4.2.9}$$

where the right-hand side is a Gauss-Hermite polynomial evaluated at a series of quadrature points indexed by q. Hedeker and Gibbons (1994) give a detailed discussion and also consider the multicategory (multinomial) response case. This function is then maximised using a suitable search procedure over the parameter

space. Estimates produced from, for example, a PQL analysis will help to locate the search region.

If we consider the model with a single random intercept at level 2 we have

$$P(u_j) = \prod_i \frac{\exp(X_{ij}\beta + u_j)}{\{1 + \exp(X_{ij}\beta + u_j)\}^2}, \quad f(u_j) = \sigma_u^2 \phi \tag{4.2.10}$$

where ϕ is the standard normal density. The standard quadrature method selects points centred around zero, but the u_j are not centred at zero and we may therefore need a very large number of quadrature points to cover the range. A solution is to use *adaptive quadrature.* One possibility (Liu and Pierce, 1994) is to centre the quadrature points on the modal (or mean) values of the function $P(u_j)$ and to scale them suitably, for example, according to the estimated standard deviation of u_j (see Hartzel *et al*., 2001 for a discussion). These central values need to be estimated and a convenient choice is to use preliminary PQL estimates together with estimated standard errors.

Quadrature methods have been applied successfully to Poisson, binomial and multinomial and ordered category models and have been implemented in software packages (SAS, MIXOR, AML, STATA (GLLAMM)). Nevertheless, successful quadrature, even with the adaptive method, will often require a large number of quadrature points and even in simple cases convergence can be difficult to achieve (Lesaffre and Spiessens, 2001). This becomes especially important when there are several random coefficients since the quadrature points will now be in several dimensions so that the number of points increases geometrically with the number of random coefficients. This places a practical limit on the complexity of models that can be handled in this way and applications have been mainly restricted to 2- or 3- level models with small numbers of random terms. Guass-Hermite quadrature is effectively limited to the normal distribution because of the exponential term in (4.2.9) and alternative quadrature methods are required for other higher level distributions.

Appendix 4.3 MCMC estimation for generalised linear models

4.3.1 Metropolis-Hastings (MH) sampling

In Appendix 2.5, the basic Gibbs sampling algorithm for models with normal random effects was described. In the case of generalised linear models, we cannot easily write down the conditional distribution for every step of the algorithm, for example for the fixed effects and the residuals. Thus, for the fixed effects step with a single level 2 variance for a binary response we have

$$p(\beta|y, \sigma_u^2, \sigma_e^2, u) \propto L(y; \beta, u, \sigma_e^2)p(\beta)$$

where

$$L(y; \beta, u, \sigma_e^2) = (1 + e^{-(X\beta+u)})^{-y}(1 + e^{-(X\beta+u)})^{y-1}$$

and a suitable diffuse prior is $p(\beta) \propto 1$

We can therefore use MH sampling (Section 2.4.2) and in general can mix Gibbs sampling with MH sampling steps. In other respects the estimation is as for normal models with the corresponding likelihood functions for binomial, Poisson, multinomial etc. data replacing the normal likelihood (see Browne and Draper, 2000, for more details).

4.3.2 Latent variable models for binary data

In the case of the probit link function for binomial data we can avoid the use of MH sampling and obtain an additional interpretation as follows. We consider a binary response where a positive response (=1) occurs when the value of an underlying continuous variable exceeds a threshold. For example, an examination may yield pass/fail grades which can be supposed to have such an underlying continuous scale. To illustrate write a variance components 2-level model for the underlying continuous variable as

$$y_{ij} = (X\beta)_{ij} + u_j + e_{ij} \tag{4.3.1}$$

and suppose a positive value occurs when $y_{ij} > 0$. We then have

$$\Pr(y_{ij} > 0) = \Pr(e_{ij} > -[(X\beta)_{ij} + u_j]) \tag{4.3.2}$$

Now if we assume $e_{ij} \sim N(0, 1)$, the probability in (4.3.2) is

$$\int_{-[(X\beta)_{ij}+u_j]}^{\infty} \phi(t)dt \tag{4.3.3}$$

where $\phi(t)$ is the pdf of N(0,1).

In the standard Gibbs algorithm, we insert an extra step which generates a random value y_{ij} from the truncated normal distribution defined by (4.3.3) given current parameter values for each level 1 unit. When the response value is a 1 we select from $[-X^*, \infty]$, $X^* = (X\beta)_{ij} + u_j$ and where the response is 0 from $[-\infty, X^*]$. These are then used with (4.3.1) as in the standard Normal case. An advantage of this approach is that it is generally faster than using MH. Additionally, however, because it generates a set of 'imputed' continuous responses we can readily combine it with observed Normal responses so allowing estimates of underlying correlations at level 1 between binary and continuous variables or between several binary variables (see Chapter 7 for details).

A similar approach can be used for the logit and log-log link functions, although these do not lead to multivariate Normal distributions where more than one response is involved and so are less useful.

For logit link models we need to take a random draw from the logistic distribution with the following density function

$$f(x) = \frac{\exp(x)}{[1 + \exp(x)]^2} \tag{4.3.4}$$

which has zero mean and variance 3.290 and where the required cumulative function is obtained as

$$\int_{-Y}^{\infty} f(x)dx = [1 + \exp(-Y)]^{-1} \tag{4.3.5}$$

which is the logit link function, so that we take a random draw as above, but this time from the truncated logistic distribution with truncation point corresponding to $X^* = (X\beta)_{ij} + u_j$ as before. To select a random draw from this logistic distribution we first take a random draw, u, from the uniform (0,1) distribution and then make the transformation $y = \log(u/(1-u))$. The corresponding 2-level model now has the level 1 variance constrained to 3.29.

For the log-log link, we take a random draw from the Gumbel distribution with density function

$$f(x) = \exp(-x)(\exp -[\exp(-x)]) \tag{4.3.6}$$

which has mean –0.577 and variance 1.645. We have

$$\int_{-Y}^{\infty} f(x)dx = 1 - \exp(-\exp(Y)) \tag{4.3.7}$$

which is the complementary log-log link function and the transformation from the uniform is now $y = \log(-\log(1-u))$ and we draw from the truncated distribution with truncation point given by $X^* = (X\beta)_{ij} + u_j$ and the level 1 variance constrained to 1.645.

Ordered responses can be fitted using an extension that involves sampling the cut point or threshold parameters and details are given in Chapter 7.

4.3.3 Proportions as responses

We can extend the (0,1) response model to the case where the response is a proportion. For the simple binomial probit model suppose there are r 'successes' out of n 'trials'. Assuming independent sampling, we generate a set of r values y_{ij} where the response is 1, i.e. from $[-X^*, \infty]$, $X^* = (X\beta)_{ij} + u_j$, and $n - r$ values from $[-\infty, X^*]$. We can think of this set of values as defining a lowest level cluster, essentially a replication level, below the original level 1 of the model and proceed to fit this modified model with an extra level where there is a single variance term constrained to be equal to 1. This leads to a simple modification to the Gibbs algorithm (Appendix 2.5). Similar modifications can be made for the other models described above.

Appendix 4.4 Bootstrap estimation for multilevel generalised linear models

4.4.1 The iterated bootstrap

As pointed out in Section 3.6, we can use the bootstrap to provide a bias correction for parameter estimates. This only works, however, if the bias resulting from a particular estimation procedure is independent of the true value of the underlying parameter. In generalised linear models this is not the case and we need to introduce a modification.

We illustrate the procedure with a simple 2-level variance components model, as follows

$$\begin{aligned} &\text{logit}(\pi_{ij}) = \beta_0 + \beta_1 x_{ij} + u_j \\ &u_j \sim N(0, \sigma_u^2) \\ &y_{ij} \sim \textit{Binomial}(1, \pi_{ij}) \end{aligned}$$

Given a set of initial estimates, obtained using for example the first order MQL approximation,

$$\hat{\sigma}_u^{2(0)}, \hat{\beta}_0^{(0)}, \hat{\beta}_1^{(0)} \tag{4.4.1}$$

we generate a set of bootstrap samples, parametrically or using the residuals bootstrap (Section 3.6) from the model using the estimates (4.4.1). Averaging over these we obtain the set of bootstrap estimates

$$\tilde{\sigma}_u^{2(0)}, \tilde{\beta}_0^{(0)}, \tilde{\beta}_1^{(0)} \tag{4.4.2}$$

We now obtain the bootstrap estimate of the bias by subtracting (4.4.2) from (4.4.1). These bias estimates are added to the initial parameter estimates (4.4.1) as a first adjustment to give new bias-corrected estimates

$$\hat{\sigma}_u^{2(1)}, \hat{\beta}_0^{(1)}, \hat{\beta}_1^{(1)} \tag{4.4.3}$$

We generate a new set of bootstrap samples from the model based upon the estimates given by (4.4.3), subtract the new mean bootstrap parameter estimates from (4.4.3) to obtain updated bias estimates and add these to the initial estimates (4.4.1) to obtain a new set of bias corrected estimates. When this process converges, Kuk (1995) demonstrates that it gives asymptotically consistent and unbiased parameter estimates.

Care needs to be taken with small variance estimates. To estimate the bias we need to allow negative estimates of variances. If an initial estimate is zero, then clearly, resetting negative bootstrap sample means to zero implies that the bias estimate will never be negative, so the new updated estimate will remain at zero. Moreover, as confirmed by simulations, all the estimates will exhibit a downward bias if negative

bootstrap means are reset to zero. We also note that where an unbiased variance estimate is close to zero, the value of the bias is anyway small, so that full bias correction is less important and, for example, a second order PQL estimate may be adequate.

The bootstrap replicates from the final bootstrap set generally will have too small a variance and so cannot directly be used for inference. If we knew the functional relationship between the bias-corrected value and the biased value, this could be used to transform each of the bootstrap replicate estimates and the transformed values then used for inference. Alternatively, for each parameter in turn, using the final bias-corrected estimate and the final bootstrap replicate mean, we take the ratio of these and multiply all the final replicate parameter values by this ratio. These scaled values can used to construct approximately correct standard errors and quantile estimates. Care is needed, however, when the initial parameter estimates are close to zero.

5

Models for repeated measures data

5.1 Repeated measures data

When measurements are repeated on the same subjects, for example students or animals, a 2-level hierarchy is established with measurement repetitions or occasions as level 1 units and subjects as level 2 units. Such data are often referred to as 'longitudinal' – as opposed to 'cross-sectional', where each subject is measured only once. Thus, we may have repeated measures of body weight on growing animals or children, repeated test scores on students or repeated interviews with survey respondents. It is important to distinguish two classes of models which use repeated measurements on the same subjects. In one, earlier measurements are treated as covariates rather than responses. This was done for the educational data analysed in Chapters 2 and 3, and usually will be appropriate when there are a small number of discrete occasions and where different measures are used at each one. (See Chapter 8 for a discussion of the extension of such models in structural equation modelling.) In the other class of model, usually referred to as 'repeated measures' models, the measurements are treated only as a set of observed responses, and it is this class of models we discuss in this chapter. A detailed description of the distinction between the former 'conditional' models and the latter 'unconditional' models can be found in Goldstein (1979) and Plewis (1985). Singer and Willett (2002) give detailed examples of applications both of repeated measures and conditional models.

We may also have repetition at higher levels of a data hierarchy. For example, we may have annual examination data on successive cohorts of 16-year-old students in a sample of schools. In this case, the school is the level 3 unit, year is the level 2 unit and student the level 1 unit. We will also look at an example where there are responses at both level 1 and level 2. It is worth pointing out that in repeated measures

Multilevel Statistical Models: 4th Edition Harvey Goldstein

models typically most of the variation is at level 2, so that the proper specification of a multilevel model for the data is of particular importance.

The link with the multivariate data models described in Chapter 6 is apparent when the occasions are fixed. Suppose we have measurements on the height of a sample of children at ages 11.0, 12.0, 13.0 and 14.0 years. We can regard this as consisting of a multivariate response vector of four responses for each child, and perform an equivalent analysis, for example relating the measurements to a polynomial function of age. This multivariate approach has traditionally been used with repeated measures data (see Grizzle and Allen, 1969). It cannot, however, deal properly with commonly found data that have arbitrary spacings or numbers of occasions and we do not consider it further, but see Bollen and Curran (2006) for elaborations of the multivariate model based on structural equation modelling.

In all the models considered so far, we have assumed that the level 1 residuals are uncorrelated. For some kinds of repeated measures data, however, this assumption will not be reasonable; we therefore investigate models which allow a serial correlation structure for these residuals.

We deal first with continuous response variables and we shall discuss repeated measures models for discrete response data later.

5.2 A 2-level repeated measures model

Consider a data set consisting of repeated measurements of the heights of a random sample of children. We can write a simple model as

$$y_{ij} = \beta_{0j} + \beta_{1j} x_{ij} + e_{ij} \tag{5.1}$$

This model assumes that height (Y) is linearly related to age (X) with each subject having their own intercept and slope so that

$$\begin{aligned} &E(\beta_{0j}) = \beta_0, \quad E(\beta_{1j}) = \beta_1 \\ &\text{var}(\beta_{0j}) = \sigma^2_{u0}, \quad \text{var}(\beta_{1j}) = \sigma^2_{u1}, \quad \text{cov}(\beta_{0j}, \beta_{1j}) = \sigma_{u01}, \quad \text{var}(e_{ij}) = \sigma^2_e \end{aligned}$$

There is no restriction on the number or spacing of ages, so that we can fit a single model to subjects who may have one or several measurements. We can clearly extend (5.1) to include further explanatory variables, measured either at the occasion level, such as time of year or state of health, or at the subject level, such as birthweight or gender. We can also extend the basic linear function in (5.1) to include higher order terms and can further model the level 1 residual so that, for example, the level 1 variance is a function of age.

We explore briefly a nonlinear model for growth measurements in Chapter 9. Such models have an important role in certain kinds of growth modelling, especially where growth approaches an asymptote, as in the approach to adult status in animals. In the following sections, we discuss the use of polynomial models which have a more

general applicability and for many applications are more flexible (see Goldstein, 1979, for further discussion), introducing examples of increasing complexity.

5.3 A polynomial model example for adolescent growth and the prediction of adult height

Our first example combines the basic 2-level repeated measures model with a multivariate model to show how a general growth prediction model can be constructed. The data consist of 436 measurements of the heights of 110 boys between the ages of 11 and 16 years together with measurements of their height as adults and estimates of their bone ages at each height measurement, based upon wrist radiographs. Age is measured around 13.0 years. (A detailed description can be found in Goldstein, 1989b.) We first write down the three basic components of the model, starting with a simple repeated measures model for height using a fifth degree polynomial.

$$y_{ij}^{(1)} = \sum_{h=0}^{5} \beta_h^{(1)} x_{ij}^h + \sum_{h=0}^{2} u_{hj}^{(1)} x_{ij}^h + e_{ij}^{(1)} \tag{5.2}$$

where the level 1 term e_{ij} may have a complex structure, for example a decreasing variance with increasing age.

The measure of bone age is already standardised, since the average bone age for boys of a given chronological age is equal to this age for the population. Thus, we model bone age using an overall constant to detect any average departure for this group, together with between-individual and within-individual variation.

$$y_{ij}^{(2)} = \beta_0^{(2)} + \sum_{h=0}^{1} u_{hj}^{(2)} x_{ij}^h + e_{ij}^{(2)} \tag{5.3}$$

For adult height, we have a simple model with an overall mean and level 2 variation. If we had more than one adult measurement on individuals, we would be able to estimate also the level 1 variation among adult height measurements; in effect, allowing for measurement errors.

$$y_j^{(3)} = \beta_0^{(3)} + u_{0j}^{(3)} \tag{5.4}$$

We now combine these into a single model, using the following indicator variables:

$\delta_{ij}^{(1)} = 1$, if growth period measurement, 0 otherwise

$\delta_{ij}^{(2)} = 1$, if bone age measurement, 0 otherwise

$\delta_j^{(3)} = 1$, if adult height measurement, 0 otherwise

$$y_{ij} = \delta_{ij}^{(1)} \left(\sum_{h=0}^{5} \beta_h^{(1)} x_{ij}^h + \sum_{h=0}^{2} u_{hj}^{(1)} x_{ij}^h + e_{ij}^{(1)} \right) + \delta_{ij}^{(2)} \left(\beta_0^{(2)} + \sum_{h=0}^{1} u_{hj}^{(2)} x_{ij}^h + e_{ij}^{(2)} \right) + \delta_j^{(3)} \left(\beta_0^{(3)} + u_{0j}^{(3)} \right)$$

At level 1 the simplest model which we shall assume is that the residuals for bone age and height are independent, although dependencies could be created, for example, if the model was incorrectly specified at level 2. Thus, level 1 variation is specified in terms of two variance terms. Although the model is strictly a multivariate model, because the level 1 random variables are independent it is unnecessary to specify a 'dummy' level 1 with no random variation as with standard multivariate models (see Chapter 6). If, however, we allow correlation between height and bone age, then we will need to specify the model with no variation at level 1, the variances and covariance between bone age and height at level 2 and the between-individual variation at level 3.

Table 5.1 shows the fixed and random parameters for this model, omitting the estimates for the between-individual variation in the quadratic coefficient of the polynomial growth curve. We see that there is a large nonzero correlation between adult height and height at age 13.0 years, and small correlations between adult height and the height growth rate and the bone age coefficients. This implies that the height and bone age measurements can be used to make predictions of adult height. For a new individual, with information available at one or more ages on height or bone age, we now show how such predictions can be obtained.

Consider the case where we have height measures at two ages, x_1, x_2. Under normality, our prediction formula will be linear and can be written as

$$\hat{y}_j^{(3)} = \beta_0^{(3)} + \alpha_1 \tilde{y}_{1j}^{(1)} + \alpha_1 \tilde{y}_{2j}^{(1)} \tag{5.5}$$

where the raw residual is

$$\tilde{y}_{ij}^{(1)} = y_{ij}^{(1)} - \sum_{h=0}^{5} \beta_h^{(1)} x_{ij}^h, \quad i = 1, 2$$

The multivariate normal covariance matrix for the three variables in (5.5) is given by

$$\begin{pmatrix} \hat{y}_j^{(3)} \\ \tilde{y}_{1j}^{(1)} \\ \tilde{y}_{2j}^{(1)} \end{pmatrix} = MVN(0, \Omega)$$

$$\Omega = \begin{pmatrix} \sigma_{u0}^{(3)^2} & & \\ \sigma_{u00}^{(3,1)} + \sigma_{u01}^{(3,1)} x_{1j} & \sigma_{u0}^{(1)^2} + \sigma_{u1}^{(1)^2} x_{1j}^2 + 2\sigma_{u01}^{(1)} x_{1j} + \sigma_e^2 & \\ \sigma_{u00}^{(3,1)} + \sigma_{u01}^{(3,1)} x_{2j} & \sigma_{u0}^{(1)^2} + \sigma_{u01}^{(1)} (x_{1j} + x_{2j}) + \sigma_{u1}^{(1)^2} x_{1j} x_{2j} & \sigma_{u0}^{(1)^2} + \sigma_{u1}^{(1)^2} x_{2j}^2 + 2\sigma_{u01}^{(1)} x_{2j} + \sigma_e^2 \end{pmatrix}$$

Table 5.1 Height (cm) for adolescent growth, bone age, and adult height for a sample of boys. Age measured about 13.0 years. Level 2 variances and covariances shown; correlations in brackets.

Parameter	Estimate (s.e.)			
Fixed				
Adult Height				
Intercept	174.4			
Group (A–B)	0.25 (0.50)			
Height				
Intercept	153.0			
Age	6.91 (0.20)			
Age^2	0.43 (0.09)			
Age^3	−0.14 (0.03)			
Age^4	−0.03 (0.01)			
Age^5	0.03 (0.03)			
Bone Age				
Intercept	0.21 (0.09)			
Random				
Level 2				
	Adult Height	Height intercept	Age	Bone Age Intercept
Adult Height	62.5			
Height intercept	49.5 (0.85)	54.5		
Age	1.11 (0.09)	1.14 (0.09)	2.5	
Bone Age Intercept				0.85
Level 1 variances				
Height	0.89			
Bone age	0.18			

so that

$$\begin{pmatrix}\hat{\alpha}_1\\ \hat{\alpha}_2\end{pmatrix} = \begin{pmatrix}\sigma_{u0}^{(1)^2} + \sigma_{u1}^{(1)^2}x_{1j}^2 + 2\sigma_{u01}^{(1)}x_{1j} + \sigma_e^2 & \\ \sigma_{u0}^{(1)^2} + \sigma_{u01}^{(1)}(x_{1j}+x_{2j}) + \sigma_{u1}^{(1)^2}x_{1j}x_{2j} & \sigma_{u0}^{(1)^2} + \sigma_{u1}^{(1)^2}x_{2j}^2 + 2\sigma_{u01}^{(1)}x_{2j} + \sigma_e^2\end{pmatrix}^{-1}\begin{pmatrix}\sigma_{u00}^{(3,1)} + \sigma_{u01}^{(3,1)}x_{1j}\\ \sigma_{u00}^{(3,1)} + \sigma_{u01}^{(3,1)}x_{2j}\end{pmatrix}$$

and we substitute these estimates into (5.5).

Table 5.2 shows the estimated standard errors associated with predictions made on the basis of varying amounts of information. It is clear that the main gain in efficiency comes with the use of height, with a smaller gain from the addition of bone age.

The method can be used for any measurements, either to be predicted or as predictors. In particular, covariates such as family size or social background can be

Table 5.2 Standard errors for adult height predictions for specified combinations of height and bone age measurements.

		Height measures (age)		
		None	11.0	11.0 12.0
Bone age measures				
None			4.3	4..2
11.0		7.9	3.9	3.8
11.0	12.0	7.9	3.7	3.7

included to improve the prediction. We can also predict other events of interest, such as the estimated age at maximum growth velocity.

Pan and Goldstein (1997) derive a procedure based on such a model for estimating norms for complex growth functions of height and weight, such as acceleration coefficients at particular ages, from whatever combination of measurements happens to be available. After initial normalisation of the data, they use a longitudinal standardising sample to establish population estimates of the growth curve polynomial coefficients covariance matrix. This is then used to derive the distribution of the required growth functions.

Pan and Goldstein (1998) extend the basic polynomial model by considering spline functions that are smoothly joining polynomials with fixed join points or 'knots'. Thus, for a set of measurements on head circumference of children from birth to 16 years, after some exploration they consider the model

$$y_{ij} = \beta_{0j} + \beta_{1j} t_{ij} + \beta_{2j} t_{ij}^2 + \beta_{3j} \log(12t_{ij} + 1) + \beta_{4j} (\xi - t_{ij})_+^3 + \beta_{5j} (t_{ij} - \xi)_+^3 + e_{ij}$$

$$\theta_+ = \begin{cases} \theta \text{ if } \theta > 0 \\ 0 \text{ if } \theta \leq 0 \end{cases}$$

with knots at 2.0 and 10.0 years. The advantage of such models is that, while there is an underlying polynomial across the whole age range, the local end relationships are further modelled by the '+' function components. This, at least in part, overcomes a disadvantage of ordinary polynomials that typically fail to provide good fits at the extremes of the time period. Note that the first + function is present up to the age of 2.0 years and the second from 10.0 years onwards. Another example of the use of such splines is given by Blatchford *et al.* (2002) in modelling educational achievement as a function of class size. (We extend this model in Chapter 16 to allow mixtures of normal and non-normal, especially discrete, responses.)

5.4 Modelling an autocorrelation structure at level 1

So far, we have assumed that the level 1 residuals are mutually uncorrelated. In many situations, however, such an assumption would be false. For growth measurements, the specification of level 2 variation serves to model a separate curve for each individual, but the between-individual variation will typically involve only a few parameters, as in the previous example. Thus, if measurements on an individual are obtained very close together in time, they will tend to have similar departures from that individual's underlying growth curve. That is, there will be 'autocorrelation' between the level 1 residuals. Examples arise from other areas such as economics where measurements on each unit, for example, an enterprise or economic system, exhibit an autocorrelation structure and where the parameters of the separate time series vary across the level 2 units.

A detailed discussion of multilevel time series models is given by Goldstein *et al.* (1994). They discuss both the discrete time case, where the measurements are made at the same set of equal intervals for all level 2 units, and the continuous time case where the time intervals can vary. We develop the continuous time model here since it is more general and flexible.

To simplify the presentation, we drop the level 1 and 2 subscripts and write a general model for the level 1 residuals:

$$\text{cov}(e_t e_{t-s}) = \sigma_e^2 f(s). \tag{5.6}$$

Thus, the covariance between two measurements depends on the variance and the time difference between the measurements. The function $f(s)$ is conveniently described by a negative exponential reflecting the common assumption that with increasing time difference the covariance tends to a fixed value, $\alpha\sigma_e^2$, and this is often assumed to be zero

$$f(s) = \alpha + \exp(-g(\beta, z, s)) \tag{5.7}$$

where β is a vector of parameters for explanatory variables z. Some choices for g are given in Table 5.3.

We can apply the methods described in Appendix 9.1 to obtain maximum likelihood estimates for these models, by writing the expansion

$$f(s, \beta, z) = \left\{1 + \sum_k \beta_{k,t} z_k g(H_t)\right\} f(H_t) - \sum_k \beta_{k,t+1} z_k g(H_t) f(H_t) \tag{5.8}$$

so that the model for the random parameters is linear. Full details are given by Goldstein *et al.* (1994). Pourahmadi (1999, 2000) considers similar models but restricted to a fixed set of discrete occasions and using a different algorithm to obtain maximum likelihood estimates. (We extend this model in Chapter 17 to consider more general correlated residual structures at different levels and describe an MCMC fitting algorithm.)

Table 5.3 Some choices for the covariance function g for level 1 residuals.

$g = \beta_0 s$	For equal intervals this is a first order autoregressive series.
$g = \beta_0 s + \beta_1(t_1 + t_2) + \beta_2(t_1^2 + t_2^2)$	For time points t_1, t_2 this implies that the variance is a quadratic function of time.
$g = \begin{cases} \beta_0 s \text{ if no replicate} \\ \beta_1 \text{ if replicate} \end{cases}$	For replicated measurements this gives an estimate of measurement reliability $\exp(-\beta_1)$.
$g = (\beta_0 + \beta_1 z_{1j} + \beta_2 z_{2ij})s$	The covariance is allowed to depend on an individual level characteristic (e.g. gender) and a time-varying characteristic (e.g. season of the year or age).
$g = \begin{cases} \beta_0 s + \beta_1 s^{-1}, & s > 0 \\ 0, & s = 0 \end{cases}$	Allows a flexible functional form, where the time intervals are not close to zero.

5.5 A growth model with autocorrelated residuals

The data for this example consist of a sample of 26 boys each measured on nine occasions between the ages of 11 and 14 years in Oxford (Harrison and Brush, (1990). The measurements were taken approximately three months apart. Table 5.4 shows the estimates from a model which assumes independent level 1 residuals with a constant variance. The model also includes a cosine term to model the seasonal variation in growth with time measured from the beginning of the year. If the seasonal component has amplitude α and phase γ we can write

$$\alpha \cos(t + \gamma) = \alpha_1 \cos(t) - \alpha_2 \sin(t)$$

In the present case the second coefficient is estimated to be very close to zero and is set to zero in the following model. This component results in an average growth difference between summer and winter estimated to be about 0.5 cm.

We now fit in Table 5.5 the model with $g = \beta_0 s$ which is the continuous time version of the first order autoregressive model.

The fixed part and level 2 estimates are little changed. The autocorrelation parameter implies that the correlation between residuals three months (0.25 years) apart is 0.19.

Finally, on this topic, there will typically need to be a trade-off between modelling more random coefficients at level 2 in order to simplify or eliminate a level 1 serial correlation structure, and modelling level 2 in a parsimonious fashion so that a relatively small number of random coefficients can be used to summarise each

Table 5.4 Height as a fourth degree polynomial on age, measured about 12.25 years. Standard errors in brackets; correlations in brackets for covariance terms.

Parameter	Estimate (s.e.)		
Fixed			
Intercept	148.9		
age	6.19 (0.35)		
age^2	2.17 (0.46)		
age^3	0.39 (0.16)		
age^4	−1.55 (0.44)		
cos (time)	−0.24 (0.07)		
Random			
level 2 covariance matrix			
	Intercept	age	age^2
Intercept	61.6 (17.1)		
age	8.0 (0.61)	2.8 (0.7)	
age^2	1.4 (0.22)	0.9 (0.67)	0.7 (0.2)
level 1 variance			
σ_e^2	0.20 (0.02)		

Table 5.5 Height as a fourth degree polynomial on age, measured about 12.25 years. Standard errors in brackets; correlations in brackets for covariance terms. Autocorrelation structure fitted for level 1 residuals.

Parameter	Estimate (s.e.)		
Fixed			
Intercept	148.9		
age	6.19 (0.35)		
age^2	2.16 (0.45)		
age^3	0.39 (0.17)		
age^4	−1.55 (0.43)		
cos (time)	−0.24 (0.07)		
Random			
Level 2 covariance matrix			
	Intercept	age	age^2
Intercept	61.5 (17.1)		
age	7.9 (0.61)	2.7 (0.7)	
age^2	1.5 (0.25)	0.9 (0.68)	0.6 (0.2)
Level 1 parameters			
σ_e^2	0.23 (0.04)		
β	6.90 (2.07)		

individual. An extreme example of the latter is given by Diggle (1988), who fits just a random intercept at level 2 and serial correlation at level 1.

5.6 Multivariate repeated measures models

We have already discussed the bivariate repeated measures model where the level 1 residuals for the two responses are independent. In the general multivariate case where correlations at level 1 are allowed, we can fit a full multivariate model by adding a further lowest level, as described in Chapter 6. For the autocorrelation model this will involve extending the models to include cross correlations. For example, for two response variables with the model of Table 5.5 we would write

$$g = \sigma_{e1}\sigma_{e2}\exp(-\beta_{12}s)$$

The special case of a repeated measures model where some or all occasions are fixed is of interest. We have already dealt with one example of this where adult height is treated separately from the other growth measurements. The same approach could be used with, for example, birth weight or length at birth. In some studies, all individuals may be measured at the same initial occasion and we can choose to treat this as a covariate rather than as a response. This might be appropriate where individuals were divided into groups for different treatments following initial measurements.

Having fitted a multivariate model, we can derive the correlation matrix between measurements for given time differences. Thus, for example, if we have a bivariate model with height and weight as responses we can compare the cross-lagged correlations, that is the correlation as a function of time interval between earlier weight and later height with the correlation between earlier height and later weight. If the latter correlation for a given time interval is substantially greater than the former, we might tentatively infer that earlier height has a greater influence on subsequent weight, in a causal sense, rather than the other way around.

5.7 Scaling across time

For some kinds of data, such as educational achievement scores, different measurements may be taken over time on the same individuals, meaning some form of standardisation may be needed before they can be modelled using the methods of this chapter. It is common to standardise such measurements so that at each measuring occasion they have the same population distribution. If this is done, we should not expect any trend in either the mean or variance over time, although generally there will be between-individual variation. An alternative standardisation procedure, when dealing with developmental measures, is to convert scores to age equivalents; that is, to assign to each score the age for which that score is the population mean or median. Where scores change smoothly with age this has the attraction of providing a readily interpretable scale. Plewis (1993) uses a variant of this in which the coefficient of

variation at each age is also fixed to a constant value. Different standardisations may be expected to lead to different inferences. The choice of standardisation is in effect a choice about the appropriate scale along which measurements can be equated so that any interpretation needs to recognise this. A further discussion of this issue is given by Plewis (1996).

We could also fit unstandardised means at each time point, or as a smoothly varying function ot time. Similarly, rather than standardise the variance prior to modelling we can choose to model it also as a function of time (see Chapter 3), and this will be important when the measurements are not made at a small number of discrete occasions.

5.8 Cross-over designs

A common procedure for comparing the effects of two different treatments A, B, is to divide the sample of subjects randomly into two groups and then to assign A to one group followed by B, and B to the other group followed by A. The potential advantage of such a design is that the between-individual variation can be removed from the treatment comparison. A basic model for such a design with two treatments, repeated measurements on individuals and a single group effect can be written as follows

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + u_{0j} + u_{2j} x_{2ij} + e_{ij} \tag{5.9}$$

where X_1 is a dummy variable for time period and X_2 is a dummy variable for treatment. In this case, we have not modelled the responses as a function of time within treatment, but this can be added in the standard fashion described in previous sections. In the random part at level 2, we allow between-individual variation for the treatment difference; we can also structure the level 1 variance to include autocorrelation or different variances for each treatment or time period.

One of the problems with such designs is so called 'carry over' effects whereby exposure to an initial treatment leaves some individuals more or less likely to respond positively to the second treatment. In other words, the u_{2j} may depend on the order in which the treatments are applied. To model this we can add an additional term to the random part of the model, say, $u_{3j}\delta_{3ij}$, where δ_{3ij} is a dummy variable which is 1 when A precedes B and the second treatment is being applied, and zero otherwise. This will also have the effect of allowing level 2 variances to depend on the ordering of treatments. Jones and Kenward (2003) give a detailed account of such designs.

5.9 Missing data

In repeated measures studies that are designed to follow up samples of individuals, it is often the case that measurements are not made at one or more target times or occasions. We have already shown that 'balanced' data are not a requirement for efficient estimates; some studies 'rotate' individuals or higher level units in or

out of the study by design. Where missing measurements on individuals or units are unscheduled, certain important issues arise. Laird (1988) provides a detailed discussion of the three situations which can be distinguished.

The first case is where missingness is completely at random (MCAR); that is, the probability of being missing, the 'non-response' probability, is independent of any of the other responses for that individual or higher level unit. This is effectively equivalent to the 'missing by design' case and all the procedures we have discussed can be applied.

The second case is where missingness is at random (MAR); that is, the non-response probability depends only on observed responses. For example, in a health interview study, previous ill health may increase the probability of non-attendance at a subsequent interview. We may also include here the possibility that we can achieve MAR by introducing covariates into the model. For example, social background may affect the probability of non-response over time. Broadly speaking, so long as the model specifying the covariance structure of the responses is correct, we can apply our previous procedures. The joint probability of the observed responses given the model parameters, is independent of the joint probability of the unobserved responses given the parameters, so that any estimation method, such as maximum likelihood or MCMC, which is based upon the joint probability function of the observations will provide valid inferences.

The third case is more troublesome, and it is where the probability of response is not independent of the unobserved, missing, values. In general, our previous procedures will lead to biases unless they are suitably modified. One approach to the problem, the 'pattern mixture' model, essentially groups individuals according to different patterns of missing responses. The patterns can have different models for the non-response, so that the response values are considered as dependent on the response probabilities. Another approach is to suppose that the response probabilities are dependent on the data, both observed and unobserved, and also are a function of time. These models are known as selection models. Diggle and Kenward (1994) develop such a model for 'monotone drop out' where individuals do not return to a study once they fail to respond at a target occasion. Kenward (1998) further develops this and introduces a sensitivity analysis to study how robust inferences are to choice of model structure. In these cases, a model for the response missingness mechanism is modelled jointly with the target model, for example, a polynomial growth curve. Such multiprocess models are discussed further in Chapter 16, where we go into more detail about the treatment of missing data.

Touloumi *et al*. (1999) give an example of a multiprocess model, using EM estimation, where changing cell counts, in a growth component, is modelled jointly with (log) patient survival time and where dropout is not independent of the survival time. In their model they do not have an explicit model for the probability of dropout, but essentially impute the unknown survival time at each iteration of the algorithm. Crouchley and Ganjali (2002) provide a general treatment of these models.

A useful specification for these models is to employ the probit link function in the response model, as described in Appendix 4.3, assuming multivariate normality for the growth component. This then allows us to estimate the covariance at the

occasion level between the 'propensity to respond' and the growth measurement. This covariance can be a function of time or time-related covariates and such models can, in principle, be fitted within an MCMC framework and are not restricted to monotone dropout structures. At the level of the individual, we allow the response to incorporate random effects which will covary with the individual random effects for the growth model, although in the monotone dropout case we will generally not be able to fit separate random effects at the occasion and individual level. The model generalises straightforwardly, using the procedures we have outlined, to multivariate growth processes. (We return to such models in Chapter 16.)

5.10 Longitudinal discrete response data

If we have repeated measures for discrete responses then a natural approach is to use the models of Chapter 4 in a 2-level model with occasions as level 1. Thus we could write for a binary response

$$\begin{aligned} &f(\pi_{ij}) = \beta_{0j} + \beta_{1j}x_{ij} \\ &E(\beta_{0j}) = \beta_0, \quad E(\beta_{1j}) = \beta_1 \\ &\text{var}(\beta_{0j}) = \sigma^2_{u0}, \quad \text{var}(\beta_{1j}) = \sigma^2_{u1}, \quad \text{cov}(\beta_{0j}, \beta_{1j}) = \sigma_{u01}, \quad \text{var}(e_{ij}) = \sigma^2_e \\ &y_{ij} \overset{iid}{\sim} Bin(1, \pi_{ij}) \end{aligned} \tag{5.10}$$

with straightforward extensions to ordered and unordered responses. In some cases such a formulation may be reasonable but often the assumption that the binomial responses are independent and identically distributed (*iid*) conditional on the π_{ij}, is unrealistic. Thus, for example, in a study of voting patterns on three occasions Yang *et al*. (2000) used a three level model for repeated binary responses of voting behaviour. They found that there was considerable under-dispersion in the data, with an extra-binomial parameter as low as 0.4, and attributed this to the fact that for many people the probability of voting for a particular party (Conservative) was effectively 0 or 1, which implies in (5.10) large numbers of individuals with infinite random effect values on the linear scale.

One method for handling such data is the so-called mover-stayer model which extends (5.10) by writing the probability of being a 'stayer', that is, not changing, as

$$\pi_{1j} = f[(X_1\alpha_1)_j + u_{1j}]$$

the probability of voting Conservative given that the individual is a stayer as

$$\pi_{2ij} = f[(X_2\alpha_2)_{ij} + u_{2j}]$$

and the probability of voting Conservative if the individual is a mover as

$$\pi_{3ij} = f[(X_3\alpha_3)_{ij} + u_{3j}]$$

For those individuals who have nonconstant responses, the probability of being a stayer is 0. We can combine these probabilities with (5.10) so that the probability of voting Conservative is

$$\pi_{1j}\pi_{2ij} + (1 - \pi_{1j})\pi_{3ij}$$

and the relevant parameters can be estimated, e.g. by maximum likelihood or MCMC.

One problem with this formulation is the assumption that some people really are true stayers and that this can describe the response pattern. In the voting example with only three occasions, this seems questionable when nearly three-quarters of the sample did not change their vote. In other cases, for example in modelling the presence of a disease over time, such an assumption may also be questionable.

To avoid relying on such a model we can directly model the responses as multivariate and estimate the covariance structure, where occasions are treated as variates. Thus, in the voting case, we would have a 3-variate model which we can fit as an eight-category multinomial response (see Chapter 4) with three dummy variable terms for occasions where these terms are associated with the average response at each occasion. We can also include covariates, possibly interacting with these dummy variables. At each occasion the coefficient of each dummy variable is assumed to vary binomially across individuals and the covariances between these are parameters to be estimated. A simple intercept model can be written

$$\text{logit}(\pi_{tij}) = \sum_{t=1}^{3} \beta_t z_{tij}, \quad y_{tij} \sim bin(1, \pi_{tij}) \tag{5.11}$$

where t indexes occasion and the z_{tij} are the dummy variables for occasions. If we use a probit link function we can directly estimate and interpret these covariances as correlations.

This multivariate formulation, however, as in the case of continuous responses, is relatively inflexible. Barbosa and Goldstein (2000) extend this model to the general case with arbitrary numbers and spacing of occasions. Essentially, they use the serial correlation models in Section 5.5, adapted for binary responses where the correlation between occasions becomes a function of the time difference. They find that the final choice of covariance function in Table 5.3 involving an inverse polynomial fits the data well, and show that this is an efficient method for handling such longitudinal data and accounting for the under-dispersion that results from a large proportion of unchanging responses. The method is readily generalised to the ordered or unordered category case, although the latter will usually lead to a large number of potential parameters. A particular advantage of this method is that it models the correlation between occasions as a smooth function of time, and possibly covariates, so allowing flexible predictions of future probabilities based on current observations to be made for individuals or groups of individuals. An MCMC algorithm for such models is described in Chapter 17.

6

Multivariate multilevel data

6.1 Introduction

In previous chapters, we basically considered only a single response variable, although we looked at special cases of multiple responses, such as in describing repeated measures data. We now look more systematically at models where we wish simultaneously to model several responses as functions of explanatory variables. We begin with the case where these are all defined at level 1, confining ourselves to the case of normal responses. (In a later section, we consider models where responses can be defined at several levels; in the next chapter, we look at multivariate models where the responses are mixtures of continuous and discrete data; and in Chapter 16, at models where we have multivariate responses of different types at several levels.)

The formulation of multivariate models given here provides us with basic tools for tackling a very wide range of problems. These problems include missing response data and rotation or matrix designs for surveys and (see Chapter 16), methods for handling very general missing data structures.

We will develop the model using a dataset of examination results consisting of scores on two components of a science examination taken in 1989 by 1905 students in 73 schools and colleges. The examination is the General Certificate of Secondary Education (GCSE) taken at the end of compulsory schooling in England, normally when students are 16 years of age. The first component is a traditional written question paper (marked out of a total score of 160) and the second consists of coursework (marked out of a total score of 108), including projects undertaken during the course and marked by each student's own teacher. The overall teachers' marks are subject to external 'moderation' using a sample of coursework. Interest in these data centres on the relationship between the component marks at both the school and student level, whether there are gender differences in this relationship and whether the variability differs for the two components. Creswell (1991) gives a full description of the dataset.

Multilevel Statistical Models: 4th Edition Harvey Goldstein

6.2 The basic 2-level multivariate model

To define a multivariate, in the case of the following example a 2-variate, model we treat the individual student as a level 2 unit and the 'within-student' measurements as level 1 units. Each level 1 measurement 'record' has a response, which is either the written paper score or the coursework score. The basic explanatory variables are a set of dummy variables that indicate which response variable is present. Further explanatory variables are defined by multiplying these dummy variables by individual level explanatory variables such as gender.

The data matrix for three individuals, two of whom have both measurements and the third with only the written paper score, is displayed in Table 6.1. The first and third students are female (1) and the second is male (0).

The model is written as

$$
\begin{aligned}
y_{ij} &= \beta_{01} z_{1ij} + \beta_{02} z_{2ij} + \beta_{11} z_{1ij} x_j + \beta_{12} z_{2ij} x_j + u_{1j} z_{1ij} + u_{2j} z_{2ij} \\
z_{1ij} &= \left\{ \begin{array}{c} 1 \text{ if written} \\ 0 \text{ if coursework} \end{array} \right\}, z_{2ij} = 1 - z_{1ij}, x_j = \left\{ \begin{array}{c} 1 \text{ if female} \\ 0 \text{ if male} \end{array} \right\} \qquad (6.1) \\
\mathrm{var}(u_{1j}) &= \sigma^2_{u1}, \quad \mathrm{var}(u_{2j}) = \sigma^2_{u2}, \quad \mathrm{cov}(u_{1j} u_{2j}) = \sigma_{u12}
\end{aligned}
$$

There are several features of this model. There is no level 1 variation specified because level 1 exists solely to define the multivariate structure. The level 2 variances and covariance are the (residual) between-student variances. In the case where only the intercept dummy variables are fitted, and every student has both scores, the model estimates of these parameters become the usual between-student estimates of the variances and covariance. The multilevel estimates are statistically efficient even where some responses, as for student 3, are missing, and where the measurements have a multivariate normal distribution they are also the maximum likelihood estimates. This formulation as a 2-level model thus allows for the efficient estimation of a covariance matrix with missing responses. Note that, while it is convenient for basic algorithmic and expository purposes to use the first subscript to denote the multivariate structure, when we come to generalise this model we will drop this notation.

Table 6.1 Data matrix for examination example (part).

		Intercepts		Gender	
Student	Response	Written	Coursework	Written	Coursework
1 (female)	y_{11}	1	0	1	0
1	y_{21}	0	1	0	1
2 (male)	y_{12}	1	0	0	0
2	y_{22}	0	1	0	0
3 (female)	y_{13}	1	0	1	0

Table 6.2 Bivariate models for written paper and coursework responses.

Fixed	Estimate (s.e.)	Estimate (s.e.)
Constant: Written	49.5	49.5
Coursework	69.5	69.1
Female: Written	–2.5 (0.5)	–2.5 (0.5)
Coursework	6.9 (0.7)	7.3 (1.1)
Random		
Level 3:		
σ^2_{v1}	48.9 (9.5)	49.6 (9.5)
σ_{v12}	25.2 (9.1)	35.5 (11.3)
σ^2_{v2}	77.1 (14.8)	106.6 (21.7)
σ_{v13}		–15.9 (7.8)
σ_{v23}		–37.4 (13.2)
σ^2_{v3}		41.5 (11.7)
Level 2:		
σ^2_{u1}	124.3 (4.1)	124.2 (4.1)
σ_{u12}	74.6 (3.9)	73.6 (3.9)
σ^2_{u2}	183.2 (6.1)	189.1 (8.6)
σ_{u23}		–12.5 (4.7)
–2 log(likelihood)	29718.8	29664.7

For the random parameters the subscripts refer to the following explanatory variables: 1 = writing intercept, 2 = coursework intercept, 3 = coursework female.

In our example, the students are grouped within examination centres, so that the centre is the level 3 unit. Table 6.2 presents the results of two models fitted to these data. The first analysis is simply (6.1) with variances and a covariance for the two components added at level 3. In the second analysis, additional variance terms for gender, statistically significant at the 5 % level, have been added at levels 2 and 3.

In both analyses, the females do worse on the written paper and better on the coursework assessment. There is a greater variability of marks on the coursework element, even though this is marked out of a smaller total, and the centre variance partition coefficients are approximately the same in the first analysis ($48.9/(48.9 + 124.3) = 0.28$ and $77.1/(77.1 + 183.2) = 0.30$). This suggests that the 'moderation' process has been successful in maintaining a similar relative between-centre variation for the coursework marks. The correlation between the two components is 0.50 at the student level and 0.41 at the centre level.

In the second analysis, we see that the between-student variance for coursework is smaller for the females ($189.1 - 2 \times 12.5 = 164.1$) compared to that for the males (189.1); for the centres, the coursework variance for females is also smaller ($106.6 - 2 \times 33.4 + 41.5 = 73.3$) than for males (106.6). There appears to be no difference in the variances for the written paper and the corresponding parameter is omitted.

Note how the standard error of the coursework gender coefficient increases with the more precise specification of the coursework variation at both levels. This is another aspect of the effect we saw when fitting a multilevel model, as opposed to a single level model.

6.3 Rotation designs

We have already pointed out that fully balanced multivariate designs are unnecessary when we formulate the model as in Table 6.1. As this shows, the basic 2-level formulation does not recognise that a response is missing, since we only record those present.[1] We now look at designs where responses are deliberately missing (as part of the design) and we see how this can be useful in a number of circumstances.

In many kinds of surveys, the amount of information required from respondents is so large that it is unrealistic to expect each one to respond to all the questions or items: in education surveys, we may require achievement information covering a large number of areas; in surveys of businesses, we may wish to have a large amount of detailed information from each business; and in household questionnaires, we may wish to obtain information on a wide range of topics. In this chapter we consider only measurements that are to be used as responses in a model (see Chapter 16 for how this can be extended to deal with explanatory variables having missing values).

If we denote the total set of responses as $\{N\}$, then we choose p subsets $\{N_i, i = 1, \ldots p\}$ each of which is suitable for administering to a level 1 unit such as a student or household. When choosing these subsets we can only estimate subject-level covariances between those responses that appear together in the same subset. It is therefore common in such designs to ensure that every possible pair of responses is present. If we wish to estimate covariances for higher level units such as schools, it is necessary only to ensure that the relevant pair of responses are assigned to some schools – a large enough number to provide efficient estimates. The subjects are assigned at random to subsets and higher level units are also assigned randomly, possibly with stratification.

Each subset is viewed formally as a multivariate response vector with randomly missing values; that is, those that are excluded from the subset. As we saw in Section 6.2, we can fit a multivariate response model for such data and obtain efficient estimates for the fixed part coefficients and covariance structures at any level. In this formulation, the variables to be used as explanatory variables should be measured for each level 1 unit. There follows an example using educational achievement data.

6.4 A rotation design example using Science Survey test scores

The data come from the Second International Science Survey carried out by the International Association for the Evaluation of Educational Achievement (Rosier,

[1] In MCMC estimation, each missing data value is treated as a parameter to be sampled, so that subsequent steps can condition on a complete dataset (as detailed in Chapter 16).

Table 6.3 Numbers of items in topic areas: Grade 8.

Form	Earth Science	Biology	Physics
1 (Core)	6	10	10
2 (R2)	–	–	7
3 (R3)	–	4	–
4 (R4)	–	4	–

1987). Table 6.3 shows how items from three science topic areas are distributed over test papers or forms with the numbers of items in each topic area. The tests consisted of a core form, or subtest, taken by all students plus a randomly selected pair out of the four additional forms or subtests. The study was carried out in 1984 in some 24 countries. We discuss here the results for Hungary.

Because the number of items in some of the additional forms for some subjects is very small, only the additional forms 2–4 are used, labelled Biology R3, Biology R4 and Physics R2 (Table 6.3). We also divide each subtest score by the total number of items in the subtest so as to reduce each score to the same scale. There are 99 schools with 2439 students and a total of 10 971 responses.

The full model for Table 6.4 can be written as a 3-level model as follows:

$$
\begin{aligned}
y_{ijk} = {} & \beta_{01}z_{1ijk} + \beta_{02}z_{2ijk} + \ldots., \beta_{06}z_{6ijk} + \beta_{11}z_{1ijk}x_{jk} + \beta_{12}z_{2ijk}x_{jk} \\
& + \ldots., \beta_{16}z_{6ijk}x_{jk} + u_{1jk}z_{1ijk} + u_{2jk}z_{2ijk} + \ldots., u_{6jk}z_{6ijk} + v_{1k}z_{1ijk} \\
& + v_{2k}z_{2ijk} + \ldots., v_{6k}z_{6ijk} \\
& \begin{pmatrix} u_{1jk} \\ u_{2jk} \\ \vdots \\ u_{6jk} \end{pmatrix} \sim N(0, \Omega_u), \begin{pmatrix} v_{1k} \\ v_{2k} \\ \vdots \\ v_{6k} \end{pmatrix} \sim N(0, \Omega_v)
\end{aligned}
$$

where subscripts $1, \ldots .6$ correspond to the six subsets in Table 6.4 We see that the intercorrelations at the student level are low and are higher at the school level. One reason for this is the fact that there are few items in each subtest so that the reliability of the tests is rather low which will decrease the correlations at student level but less so at school level. Because of these low correlations among the subtests the joint analysis does not result in a marked improvement in efficiency when we compare this analysis with an analysis for a single subtest. For example, if we fit a univariate model for the Physics R2 subtest using the 1226 students responding to that subtest, we obtain fixed part estimates of 0.665 (0.0132) and –0.073 (0.0124), which are close to those above and with standard errors only slightly higher.

In order to provide the most precise estimates we treated the subtests separately, although we would generally wish to make inferences for each subject area, combining over the tests. The natural way to do this is to form a weighted average of the subtest estimates, in this case weighting by the number of items in each subtest.

Table 6.4 Science attainment estimates for Hungary IEA study.

Fixed	Estimate (s.e.)
Earth Science Core	0.838 (0.0076)
Biology Core	0.711 (0.0100)
Biology R3	0.684 (0.0109)
Biology R4	0.591 (0.0167)
Physics Core	0.752 (0.0128)
Physics R2	0.664 (0.0128)
Earth Science Core (girls-boys)	−0.0030 (0.0059)
Biology Core (girls-boys)	−0.0151 (0.0066)
Biology R3 (girls-boys)	0.0040 (0.0125)
Biology R4 (girls-boys)	−0.0492 (0.0137)
Physics Core (girls-boys)	−0.0696 (0.0073)
Physics R2 (girls-boys)	−0.0696 (0.0116)

Random Variances on diagonal; correlations off-diagonal

School level

	E.Sc. core	Biol. core	Biol R3	Biol R4	Phys. core	Phys. R2
E.Sc. core	0.0041					
Biol. core	0.68	0.0076				
Biol R3	0.51	0.68	0.0037			
Biol R4	0.46	0.68	0.45	0.0183		
Phys. core	0.57	0.90	0.76	0.63	0.0104	
Phys. R2	0.54	0.78	0.57	0.65	0.78	0.0095

Student level

	E.Sc. core	Biol. core	Biol R3	Biol R4	Phys. core	Phys. R2
E.Sc. core	0.0206					
Biol. core	0.27	0.0261				
Biol R3	0.12	0.13	0.0478			
Biol R4	0.14	0.27	0.20	0.0585		
Phys. core	0.26	0.42	0.11	0.27	0.0314	
Phys. R2	0.22	0.33	0.14	0.37	0.41	0.0449

Thus, for the biology core and subtests we would form the weighted sum with weights 0.556, 0.222 and 0.222 respectively. This gives biology estimates for the boys and (girls-boys) of 0.68 (0.009) and –0.02 (0.007).

We can compare this with simply using the original scores and forming the weighted combination of the core and two subtests, eliminating any students with missing data. This results in only 399 students with complete data and the corresponding estimates are 0.68 (0.013) and –0.008 (0.015). In this case, even though the individual level 1 correlations are relatively small, the gain in efficiency from carrying out the full multivariate analysis is substantial, especially for inferences about the gender difference which in the second analysis is less than its standard error.

Another way to combine the subtests would be to form, for each student, a score based upon the items which the student responded to. Thus, for Biology the 399 students taking the core and both rotated forms would have a score out of 18 items; and there would be 823 and 807 students respectively with scores out of 14 items, with 410 students having only a score out of the core test. Since the scores are out of different totals, we would expect the between student and between-school variances to differ and this is the case; the between-student variance for the 10 core test score is 0.00013 compared to that for the 18-item core and two rotated forms score of 0.00021. In general, we would thus need to fit separate variance and covariance terms for each combination, and in effect treat the four combinations as separate responses in order to obtain efficient estimates. We would also tend to obtain high correlations between these combination scores that could lead to numerical estimation problems, so in general such a procedure is not recommended.

6.5 Informative response selection: subject choice in examinations

In the previous example, the choice of response was independent of the value of the response measurement, since allocation of students was made at random. In some cases, however, this will not be true. We now look at one such case where students taking an examination can make a choice of papers, and this choice is related to their underlying achievement.

Yang *et al*. (2001) studied a large group of students taking mathematics examinations at Advanced level in England in 1997, aged 18. There were 52 587 students in 2592 educational institutions. The data were analysed using a scoring system for achieved grades whereby the highest grade (A) is given a score of 10 and a failure given a score of 0, with a mean overall of 6.4. Students take up to four papers chosen from 10 different types with just under 87 % choosing just one paper and just under 13 % choosing two. It is clear from the data that students taking particular combinations of subjects tend to perform better than those taking other combinations. For example, the average score on the main, basic mathematics paper is 5.54 for those who take this paper alone compared to 9.45 for those who take it in combination with a further mathematics paper. The analysis adjusts for prior achievement measured by

the performance on GCSE examination taken two years earlier and interest focuses on differences between institution types and on variability and correlations among papers at the school and student level.

In order to take account of informative choice, the authors fit a term (dummy variable) for each combination of subjects taken, allow the coefficients of these to vary randomly at the school level and for the student level variance to depend on the combination chosen (see Chapter 3). They also allow interactions with other explanatory variables, such as gender. They are able to conclude that the institutional level correlations among subjects, when different choice combinations are modelled, are typically only moderate, and that institutions are more homogeneous with respect to results from some combinations as opposed to others. Models of this kind may be useful in other situations, for example, where question choice within examinations is permitted.

6.6 Multivariate structures at higher levels and future predictions

To illustrate the use of models for higher level multivariate data, we use data on school examination results for the same set of schools over two years, where there is a different set of students at each time point. Note that the responses are still at level 1; it is the multivariate structure that is at level 2.

The data are normalised examination scores taken from the national pupil database (NPD), a census of all pupils in the English state education system. Full details can be found in Leckie and Goldstein (2009). The examination results, taken in 2002 and 2007, are a subset of those available; a total of 91 453 students took their GCSE in school year 11 in a sample of 266 matched schools. There is also data on prior achievement, as measured by Key Stage 2 (KS2) test scores taken in year 6, just before entry to Secondary school. The NPD holds data on pupils' test score histories and a limited number of pupil level characteristics.

We have a bivariate model that we write in a form essentially equivalent to (6.1), where now each response defines a line of the model as follows:

$$\begin{aligned} y_{1ij} &= \beta_{10} + \beta_{11}x_{1ij} + u_{1j} + e_{1ij} \\ y_{2ij} &= \beta_{20} + \beta_{21}x_{2ij} + u_{2j} + e_{2ij} \\ \begin{pmatrix} u_{1j} \\ u_{2j} \end{pmatrix} &\sim N(0, \Omega_u), \ \Omega_u = \begin{pmatrix} \sigma^2_{u1} & \\ \sigma_{u12} & \sigma^2_{u2} \end{pmatrix} \\ \begin{pmatrix} e_{1ij} \\ e_{2ij} \end{pmatrix} &\sim N(0, \Omega_e), \ \Omega_e = \begin{pmatrix} \sigma^2_{e1} & \\ 0 & \sigma^2_{e2} \end{pmatrix} \end{aligned} \tag{6.2}$$

where the subscript 1 refers to 2002 and 2 to 2007 respectively, and x_{1ij}, $x_{2ij}$are the KS2 test scores relevant to 2002 and 2007. Since the students are not the same in the two years, there is no covariance at level 1 – but there is at level 2, since we have the same schools. In Table 6.5, we present an extract of results from the more extensive analysis in Leckie and Goldstein (2009).

Table 6.5 Selected parameter estimates for the bivariate 2-level random intercepts model of the normalised GCSE score for the 2002 and 2007 cohorts of pupils. Standard errors in brackets.

	2002		2007	
Fixed part				
Constant	−0.055	(0.014)	−0.071	(0.014)
Average KS2 score	0.667	(0.006)	0.680	(0.005)
Average KS2 (squared)	0.028	(0.003)	0.042	(0.003)
Average KS2 (cubed)	−0.026	(0.002)	−0.026	(0.001)
Female	0.189	(0.006)	0.184	(0.006)
Random Part: School				
Between-school variance (2002)	0.047	(0.004)		
Between-school covariance (2002, 2007)	0.030	(0.004)		
Between-school variance (2007)	0.047	(0.004)		
Random Part: Pupil				
Within-school between-pupil variance (2002)	0.368	(0.003)		
Within-school between-pupil variance (2007)	0.397	(0.003)		
Deviance (−2*log likelihood)	173243			
Number of schools	266			
Number of pupils	91453			

The fixed part parameter estimates for 2002 have the same signs and similar magnitudes to those for 2007. In the random part of the model, the between-school variances for the 2002 and 2007 cohorts are constrained to equal one another. A likelihood ratio test shows that this constraint, which simplifies the formula for predicting future school effects, does not significantly reduce the fit of the model ($\chi^2_{(1)} = 0.156$, $P = 0.69$). The model gives a VPC of 0.113 and 0.106 for 2002 and 2007 respectively; schools are no more or less important a source of variation in unexplained progress in 2007 than they are in 2002. The correlation between the 2002 and 2007 school effects is 0.64 (= 0.030/0.047).

We can also use these results as a basis to predict the school 'effects' (estimated residuals) five years into the future, given current data. In other words, assuming that these results can be extrapolated, given estimates for the relationship between school results five years apart, these can be used in a prediction formula. Such an analysis may haves implications, for example, for parents choosing schools, where interest is on future examination results. Thus we require

$\hat{u}_{2j} = E(u_{2j}|y_{1ij};\theta)$, where θ is the set of parameters estimated from the model (6.2). If we assume $\sigma^2_{u1} = \sigma^2_{u2} = \sigma^2_u$, Leckie and Goldstein (2009) show that the

estimates of u_{2j}, and their comparative variances, are given by

$$\hat{u}_{2j} = \frac{\rho_{u12} n_{1j} \sigma_u^2}{(n_{1j}\sigma_u^2 + \sigma_{e1}^2)} \tilde{y}_{1j}, \ \text{var}(\hat{u}_{2j} - u_{2j}) = \frac{n_{1j}\sigma_u^4(1 - \rho_{u12}^2) + \sigma_u^2 \sigma_{e1}^2}{(n_{1j}\sigma_u^2 + \sigma_{e1}^2)} \tag{6.3}$$

where $\rho_{u12} = \sigma_{u12}/\sigma_u^2$ and $\tilde{y}_{1j}$ is the mean of the raw residuals for the j-th school (see Section 2.6).

The expressions in (6.3) can be used to construct estimates and confidence intervals for each school. We see from (6.3) that predicting the time 2 estimates from those at time 1 we have a further shrinkage factor ρ_{u12} and there is a wider confidence interval than that for the time 1 residuals based solely on the time 1 data. In the present case, the confidence intervals are about 3.5 times as wide and almost all of them include zero. On the basis of these results the use of such data for school choice would be very limited and Leckie and Goldstein (2010) show that little extra precision is obtained when the prediction is based upon more than one past year's results.

6.7 Multivariate responses at several levels

We have already seen an example (Chapter 5) of a model where we simultaneously modelled level 1 and level 2 responses. We now set out a generalisation of this for multivariate normal responses. We restrict the development to a 2-level model. We go on to show in Chapter 12 how cross classifications can be introduced into multivariate models and use the results derived in Chapter 7 to extend this model to include responses other than normally distributed ones (see Chapter 16).

Let $j = 1,\ldots m$ index level 2 units and $i = 1,\ldots n_j$ index the level 1 units within the level 2 units. The underlying multivariate normal model structure we consider is as follows:

$$\begin{aligned} y_{ij}^{(1)} &= X_{1ij}\beta^{(1)} + Z_{1ij}u_j^{(1)} + e_{ij}^{(1)} \\ y_j^{(2)} &= X_{2j}\beta^{(2)} + Z_{2j}u_j^{(2)} \\ e_{ij}^{(1)} &\sim MVN(0, \Omega_1), u_j = (u_j^{(1)}, u_j^{(2)})^T, u_j \sim MVN(0, \Omega_2) \end{aligned} \tag{6.4}$$

The superscripts now denote the level at which a variable is measured. We no longer use the notation introduced earlier, where the first subscript denotes a lowest level that has no variation and simply describes the multivariate structure. Thus $y_{ij}^{(1)}$ is now a row vector that contains p_1 responses that are defined at level 1 and $y_j^{(2)}$ contains p_2 responses that are defined at level 2. Without loss of generality, we assume the same set of explanatory variables for each response at level 1 and likewise at level 2: we can constrain some of these for particular responses to be zero, as described in Appendix 6.1. The row vectorX_{1ij} contains f_1 level 1 explanatory variables and $\beta^{(1)}$ is the $(f_1 \times p_1)$ matrix containing the fixed coefficients for these. Similarly, we have X_{2j}, $\beta^{(2)}$ for the level 2 responses. For each level 1 response we have q_1 random coefficients, and Z_{1ij} is the matrix that contains the explanatory variables for these,

denoted by $u_j^{(1)}$; for each level 2 unit there are $(q_1 \times p_1)$ of these. The level 1 residuals are denoted by $e_{ij}^{(1)}$. The explanatory variable vector Z_{2j} for the level 2 residuals $u_j^{(2)}$ associated with the level 2 responses, has p_2 elements. Note that for the level 1 and level 2 responses we assume simple residual variation, but in principle we can extend our model to include complex variance structures.

Our estimation strategy uses MCMC via Gibbs sampling, drawing from the appropriate known posterior conditional distributions, and assumes appropriate diffuse prior distributions. We now summarise the steps involved (these are set out in detail in Appendix 6.1).

Step 1 Sample the level 1 fixed coefficients from the first line of (6.4) given the current parameter estimates.

Step 2 Sample the level 2 fixed coefficients in similar manner from the second line of (6.4).

Step 3 Sample the level 2 random effects $u_j^{(1)}$for the level 1 responses ignoring level 2 residuals for level 2 responses.

Step 4 Calculate the level 1 residuals $e_{ij}^{(1)}$ by subtraction.

Step 5 Calculate the level 2 residuals $u_j^{(2)}$ for the level 2 responses by subtraction.

Step 6 Sample the level 1 covariance matrix.

Step 7 Sample the full level 2 covariance matrix.

The level 2 covariance matrix sampled in Step 7 links together the level 1 and level 2 random effects in the model. Where any response is missing, we fill in a value by sampling conditional on the non-missing residuals associated with that level.

For further levels and classifications we have a similar series of steps where the covariance matrices at levels 2, 3, ... include all the random effects defined at those levels.

6.7.1 Fitting responses at several levels using random data augmentation

Many software packages are not able to implement the procedure we have just outlined. The following alternative method can be implemented in standard multilevel software.

Consider (6.4), where we now augment the level 2 responses by adding independently distributed 'noise'

$$\begin{aligned} y_{ij}^{(1)} &= X_{1ij}\beta^{(1)} + Z_{1ij}u_j^{(1)} + e_{ij}^{(1)} \\ y_{ij}^{(2)} &= X_{2j}\beta^{(2)} + Z_{2j}u_j^{(2)} + e_{ij}^{(2)} \\ e_{ij}^{(1)} &\sim MVN(0, \Omega_1),\ e_{ij}^{(2)} \overset{iid}{\sim} N(0, \sigma_{e2}^2) \\ \mathrm{u_j} &= (u_j^{(1)}, u_j^{(2)}),\ u_j \sim MVN(0.\Omega_2) \end{aligned} \tag{6.5}$$

The term $e_{ij}^{(2)}$ represents random independent level 1 noise with a predetermined small variance, σ_{e2}^2, added. This transforms the model into one where all responses are at level 1 and can be fitted in the standard way. When fitting the model, if we constrain $\text{var}(e_{ij}^{(2)})$ to equal σ_{e2}^2, then we will obtain unbiased estimators for (6.4). This then allows us to carry out, for example, imputations as described in Chapter 16. We can also use this device with more complex structures such as cross classifications.

6.8 Principal components analysis

We have already seen in Section 6.1 that the covariance matrix for a multivariate response vector where there are missing data can be efficiently estimated by arranging for the multivariate structure to constitute a 'dummy' level 1. When the variables have a multivariate normal distribution, the resulting estimates are maximum likelihood or restricted maximum likelihood.

The aim of principal components analysis is to find a linear function of a set of variates which has the maximum variance, subject to a suitable constraint. In the single level case we require to maximise the variance of $\mathbf{w}^T\mathbf{y}$, known as the first principal component, where $\mathbf{w}$ is the vector of weights defining the linear function of the variates $\mathbf{y}$, and Ω is the covariance matrix of $\mathbf{y}$, namely

$$\Lambda = \mathbf{w}^T\Omega\mathbf{w}, \quad \mathbf{w}^T\mathbf{w} = 1.$$

The solution is given by the eigenvector associated with the largest eigenvalue of Ω, that is the solution of

$$|\Omega - \lambda I| = 0$$

We define a second linear function by the set of weights that maximises the variance subject to the function being uncorrelated with the first function. The solution is given by the eigenvector associated with the second largest eigenvalue, and defines the second principal component. Subsequent functions can be defined similarly (Lawley and Maxwell, 1971). The variates are usually standardised to have equal variances.

We note that the covariance (or correlation) matrix Ω can be a residual matrix, after regressing on explanatory variables. Thus, if we wish to form a principal component for the four science subjects of the previous section, we might use the residual covariance matrix, after adjusting for gender differences. We now, however, have a choice of two covariance matrices: between-student and between-school. If we choose the between-student matrix, then we would interpret the principal component as that which had been adjusted for school differences. In forming the derived summary variable(s), we would not use the actual observed variates but the level 1 estimates of them; that is, the level 1 residuals of (6.1).

We could also choose to summarise the level 3 covariance matrix, and in this case we would use the school level residuals as the variates in the linear function. If the principal component analysis has been carried out on the residuals from a multivariate

Table 6.6 Principal Component weights for science test scores and percentage variation accounted for.

Subject	Between-student	Between-school
Earth Science Core	0.17	0.21
Biology Core	0.29	0.40
Biology R3	0.31	0.21
Biology R4	0.63	0.59
Physics Core	0.35	0.46
Physics R2	0.52	0.43
% variation	41 %	72 %

multilevel analysis, then we may wish to regard the school level principal component as a convenient summary measure of school differences.

Table 6.6 shows the student level and school level principal component weights for the Science data. Since the measures are designed to be on the same scale we work directly with the covariance matrices.

As might be expected, the components both have positive weights. At the school level, the percentage variation accounted for by the first component is high suggesting that school Science performance may usefully be summarised by this weighted function of the individual school level subject residuals. Also, the two sets of weights are fairly similar. This suggests that if we wished to summarise the individual subject scores into a single index, we could do this using the student level weights, or even the weights obtained using the total covariance matrix. In Chapter 8, we further explore the structure of these data using a factor analysis model.

6.9 Multiple discriminant analysis

Given a set of variates, we can seek a linear function of them that best discriminates among groups. This leads to the following definition. If $\overline{\mathbf{y}}$ is the vector of group means then we require a set of weights $\mathbf{w}$ such that $\mathbf{w}^T\overline{\mathbf{y}}$ has maximum variance, subject to the within-group variance of $\mathbf{w}^T\mathbf{y}$ being constrained, for example equal to 1.0. The solution is the vector associated with the largest root of

$$|\Omega_B - \lambda\Omega_W| = 0$$

for the between-group (Ω_B) and within-group (Ω_W) covariance matrices. For just two groups this gives the usual 'Fisher' discriminant function. As in principal components analysis, we can find further vectors that discriminate best, subject to being uncorrelated with all the previous vectors. The function of the variates $\mathbf{w}^T\mathbf{y}$ can then be used, for example, to classify a new unit into the 'nearest' group.

In the 2-level case our groups are the level 2 units so that we require the covariance matrices from both levels. Using the Science data example the first vector is given by the weights 0.41 –0.07 1.00 0.26 0.31 0.13 and explains about 48 % of the variation. The next two vectors account for 19 % and 13 %. It is difficult to interpret these weights and in this case the function would seem to have limited usefulness for discriminating between schools.

Appendix 6.1 MCMC algorithm for a multivariate normal response model with constraints

Consider the two level multivariate response model with p_1 level 1 responses

$$\begin{aligned} y_{ij} &= X_{1ij}\beta + z_{1ij}u_j + e_{ij} \\ e_{ij} &\sim MVN(0, \Omega_1), u_j \sim MVN(0, \Omega_2) \end{aligned} \tag{6.1.1}$$

where, for simplicity, we omit the superscript (1) defined in Section 6.7. To sample β we assume a uniform prior and sample from the posterior distribution which is multivariate normal with mean

$$\left[\sum_{ij}(I_{(p_1\times p_1)}\otimes X_{ij})^T\Omega_1^{-1}(I_{(p_1\times p_1)}\otimes X_{ij})\right]^{-1}\sum_{ij}(I_{(p_1\times p_1)}\otimes X_{ij})^T\Omega_1^{-1}\tilde{y}_{ij}^T,$$
$$\tilde{y}_{ij} = y_{ij} - z_{ij}u_j$$

and covariance matrix $[\sum_{ij}(I_{(p_1\times p_1)}\otimes X_{ij})^T\Omega_1^{-1}(I_{(p_1\times p_1)}\otimes X_{ij})]^{-1}$.

We sample u_j from the multivariate normal distribution

$$MVN\left(\left[\sum_i z_{ij}^T\Omega_1^{-1}z_{ij} + \Omega_2^{-1}\right]^{-1}\left[\sum_i z_{ij}^T\Omega_1^{-1}(y_{ij} - X_{ij}\beta)\right], \left[\sum_i z_{ij}^T\Omega_1^{-1}z_{ij} + \Omega_2^{-1}\right]^{-1}\right)$$

We sample a new level 2 covariance matrix from its posterior distribution

$$\Omega_2^{-1} \sim Wishart(v_u, S_u),\ v_u = m + v_p,\ S_u = \left(\sum_{j=1}^{m} u_j u_j^T + S_p\right)^{-1}$$

Where m is the number of level 2 units, u_j is the row vector of residuals for the j-th level 2 unit and the prior $p(\Omega_2^{-1}) \sim Wishart(v_p, S_p)$, where v_u are the degrees of freedom – the sum of the number of level 2 units and degrees of freedom associated with the prior. One choice is $v_p = -3,\ S_p = 0$ which is equivalent to choosing a uniform prior for Ω_2.

The level 1 covariance matrix is sampled in the same way. For our present model this is assumed to have only one random effect (residual) for each response. The level 1 residuals are obtained by subtraction.

Where any responses are missing we fill these in by sampling new responses, omitting detailed subscripts, by drawing from $MVN(X_2^*\beta_2^* + e_1^*\beta_1^* + z_2^*u_2^*, \Sigma_2 - \Sigma_{21}\Sigma_1^{-1}\Sigma_{12})$ where Σ_2 is the current covariance matrix of residuals for the missing

responses, Σ_1 is the covariance matrix of residuals for the observed (non-missing) responses and Σ_{12} is the matrix of covariances between the observed and missing residuals. The $X_2^*\beta_2^*$ and $z_2^*u_2^*$ are the fixed predictor and level 2 residual contribution for the missing responses, $\beta_1^* = \Sigma_1^{-1}\Sigma_{12}$ and e_1^* are the level 1 residuals for the observed responses.

Where we have responses at additional levels the sampling follows the same pattern. Thus, suppose that we have responses at level 2. We sample the fixed effects using the level 2 covariance matrix associated with the level 2 responses. The level 2 residuals for the level 2 responses are obtained by subtraction and the full level 2 covariance matrix, for responses at both levels, is sampled in a single step using the level 2 residuals for both the level 1 and level 2 responses. Where there are missing values for level 2 responses we condition on the full level 2 covariance matrix and associated residuals that are non-missing.

6.1.1 Constraints among parameters

In some cases we may wish to impose a constraint on a subset of the parameters. For example, if we wish to have a different set of predictors for each response, we can fit a maximal model and constrain appropriate subsets to zero.

Consider first, linear constraints for the fixed coefficients. Suppose we wish to impose a set of q independent linear constraints on some or all of the elements of β. These constraints will involve $q^* \geq q$ distinct elements of β, and we re-order β so that q of these q^* elements appear first.

We can write the set of constraints as

$$C\beta = k, \;\; C \text{ is } (q \times p), k \text{ is } (q \times 1) \tag{6.1.2}$$

Write the QR decomposition of C as $C = QR, R = (TW)$, where Q is orthogonal $(q \times q)$ and T is upper triangular $(q \times q)$ and W is $(q \times (p-q))$. Write $\beta^T = (\beta_1^T \beta_2^T)$ where β_1 contains the first q elements of (the re-ordered) $\beta$and β_2 contains the last p-q elements.

We now have $Q^T k = Q^T QR\beta = (TW)\beta$ and we construct

$$\begin{pmatrix} T & W \\ 0 & I \end{pmatrix} \beta = \begin{pmatrix} (TW)\beta \\ \beta_2 \end{pmatrix} = \begin{pmatrix} Q^T k \\ \beta_2 \end{pmatrix}$$

which gives

$$\begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = \beta = \begin{pmatrix} T & W \\ 0 & I \end{pmatrix}^{-1} \begin{pmatrix} Q^T k \\ \beta_2 \end{pmatrix} = \begin{pmatrix} T^{-1} & -T^{-1}W \\ 0 & I \end{pmatrix} \begin{pmatrix} Q^T k \\ \beta_2 \end{pmatrix} = \begin{pmatrix} T^{-1}Q^T k - T^{-1}W\beta_2 \\ \beta_2 \end{pmatrix}$$

so that we can sample the p-q elements of β_2 freely and compute the q elements of β_1 as the function of β_2 given by the expression $T^{-1}Q^T k - T^{-1}W\beta_2$.

We can impose a set of linear constraints in similar fashion on any subset of the random parameters.

Where we have nonlinear constraints or inequality constraints, it becomes more difficult to find a routine procedure that will separate parameters into a group that can be freely estimated and the remainder that are non-stochastically dependent on these. Nevertheless, if we do have q such constraints, and they can be separated into two such groups then we can proceed analogously to the case of linear constraints by estimating the free parameters and then each of the related parameters in turn.

Where there is an inequality constraint for a parameter of the form $k_1 \leq \theta \leq k_2$ we can use a MH proposal distribution for the parameter, where the admissible interval defines truncation points, and a suitable prior distribution for the parameter is required. See Section 3.1 for a simple example.

7

Latent normal models for multivariate data

7.1 The normal multilevel multivariate model

In the last chapter, we introduced the multilevel multivariate normal model with responses at several levels. In this chapter, we shall consider only responses at level 1 and generalise it to the case where responses can be mixtures of different types. We bring these together in Chapter 16 by considering models where there are different types of responses at different levels.

We showed in Chapter 6 how to set up a multivariate model as a 2-level model where the lowest level described the multivariate structure and the variances and covariances of the responses were defined at level 2. This provided a convenient way to obtain maximum likelihood estimates using the algorithms described in Chapter 2, but is less convenient for more complex models where we shall be using MCMC estimation. Instead, we write the basic multivariate normal model, as described in Appendix 6.1

$$\begin{aligned} y_{ij} &= X_{ij}\beta + z_{ij}u_j + e_{ij} \\ e_{ij} &\sim MVN(0, \Omega_1), \quad u_j \sim MVN(0, \Omega_2) \end{aligned} \tag{7.1}$$

Now y_{ij} is a row vector that contains the p responses.

We also described in Chapter 6 the steps needed to obtain estimates for the more general model with responses at several levels. For (7.1), the MCMC estimation steps (see Appendix 6.1 for details) are:

Step 1 Sample the level 1 fixed coefficients from the first line of (7.1) given the current parameter estimates.

Step 2 Sample the level 2 random effects u_j.

Multilevel Statistical Models: 4th Edition Harvey Goldstein

Step 3 Calculate the level 1 residuals e_{ij} by subtraction.

Step 4 Sample the level 1 covariance matrix.

Step 5 Sample the level 2 covariance matrix.

Where any response is missing we fill in a value by sampling conditional on the non-missing responses.

The following sections in this chapter show how responses of various different types can be incorporated within a multivariate normal framework. We refer to this as a 'latent normal model', where the observed non-normal responses plus the observed normal responses are mapped onto an underlying multivariate normal distribution, and describe the sampling steps to do this. In all cases, when we sample the underlying normal variables, given observed non-normal variables, we will condition on all the remaining (correlated) latent and observed normal responses. We start by considering binary responses.

7.2 Sampling binary responses

Although we can consider a binary response as a special case of a multicategory response, we consider it first in order to link with the univariate latent normal model introduced in Section 4.8, where we defined the binary response in terms of a normal threshold model.

For a given binary response, consider the corresponding latent normal variable $y_{ij,1}$, which without loss of generality we can label variable 1, and denote the remaining responses by y^*_{ij} where these are assumed to have a multivariate normal distribution either because they are observed as such or because they result from sampling the non-normal responses. We wish to sample a latent normal value for $y_{ij,1}$. Given the current parameter values, we have the conditional distribution

$$\begin{aligned} &y_{ij,1}|y^*_{ij} \sim N(X_{ij}\beta_1 + y^*_{ij}\beta_2 + z_{ij}u_j, 1 - \Sigma_{21}\Sigma_1^{-1}\Sigma_{12}) \\ &\beta_2 = \Sigma_{21}\Sigma_1^{-1}, \quad \text{cov}\begin{pmatrix} y^*_{ij} \\ y_{ij,1} \end{pmatrix} = \Sigma = \begin{pmatrix} \Sigma_1 & \\ \Sigma_{12} & \Sigma_2 \end{pmatrix}, \quad \Sigma_2 = 1 \end{aligned} \tag{7.2}$$

Thus, assuming that a 1 occurs when $y_{ij,1} > 0$, then for an observed value of 1 we sample from the normal distribution $N(0, 1 - \Sigma_{21}\Sigma_1^{-1}\Sigma_{12})$ in the interval $(-(X_{ij}\beta_1 + y^*_{ij}\beta_2 + z_{ij}u_j), \infty)$, otherwise sample from $(-\infty, -(X_{ij}\beta_1 + y^*_{ij}\beta_2 + z_{ij}u_j))$. This is similar to the procedure described in Appendix 4.3, but with the additional conditioning on the remaining correlated variable values.

7.3 Sampling ordered categorical responses

Suppose we have an ordered p-category response, with ordered categories numbered 1,. . ..p. We adopt the same conditional model as above but also introduce additional

category thresholds or cut points, $\{\alpha_k,\ k = 1, \ldots, p\}$. We have

$$y_{ij,1}|y^*_{ij} \sim N(X_{ij}\beta_1 + y^*_{ij}\beta_2 + z_{ij}u_j,\ 1 - \Sigma_{21}\Sigma_1^{-1}\Sigma_{12})$$

where $\beta_2 = \Sigma_{21}\Sigma_1^{-1}$. If we observe category $k(1 < k < p)$ then we sample from the normal interval $(\alpha_{k-1} - (X_{ij}\beta_1 + y^*_{ij}\beta_2 + z_{ij}u_j), \alpha_k - (X_{ij}\beta_1 + y^*_{ij}\beta_2 + z_{ij}u_j))$ with associated probability $\pi_{\alpha,k}$.

If $k = 0$ we sample from $(-\infty, -(X_{ij}\beta_1 + y^*_{ij}\beta_2 + z_{ij}u_j))$ and if $k = p$ we sample from $(\alpha_{p-1} - (X_{ij}\beta_1 + y^*_{ij}\beta_2 + z_{ij}u_j), \infty)$.

The α threshold parameters define the cumulative probability distribution for the ordered response. Typically, we set $\alpha_1 = 0$ so that the lowest threshold is absorbed into the intercept. They are most efficiently estimated using MH sampling, as follows.

Given the current parameters, the component of the likelihood associated with a particular ordered category response is given by

$$P_\alpha = \prod_{j=1}^{J}\prod_{i=1}^{n_j}\prod_{k=1}^{p}\pi_{\alpha,k}^{w_{ij,k}}$$

for given α where $w_{ij,k} = 1$ *iff* response for unit ij is in category k. The probabilities $\pi_{\alpha,k}$ are those corresponding to the normal intervals defined above.

We select a new set of values α^* (one at a time) using a suitable (for example normal) proposal distribution and set new threshold parameters $= \alpha^*$ with probability $\min(1, P_{\alpha^*}/P_\alpha)$. The choice of proposal distribution variance may be derived adaptively but in practice the value $5.8/N$ may be suitable (Raftery and Lewis, 1992), where N is the number of units at level 1. We also require that these threshold parameters are strictly ordered and if a proposed value violates this requirement it is not accepted. In Chapter 11, we discuss a model where this requirement is automatically satisfied via a nonlinear formulation for the thresholds.

We saw in Chapter 3 how to model the level 1 variance in terms of further variables and we can carry out an analogous procedure by modelling the threshold parameters as functions of further variables or by introducing a variance scaling factor for the level 1 variance that is a function of further variables. We discuss this in detail in Chapter 11.

We note that we cannot always sample values to achieve exact multivariate normality when the number of categories is greater than 2, since in general we have insufficient parameters to characterise the full multivariate normal distribution. Thus, for example, if we have two ordered variables each with three categories there are eight probabilities to be modelled in the (3×3) table formed by these, but we have only five parameters, namely two thresholds, two intercepts and a correlation. The linear conditional sampling procedure, however, can be expected to provide a good approximation in the sense that if an underlying latent multivariate normal distribution does exist, this sampling procedure will estimate it. We consider below a procedure for handling cases that involve partial ordering.

7.4 Sampling unordered categorical responses

We use a 'maximum indicant' model (Aitchison and Bennett, 1970), as follows. Suppose that we have just a multinomial (unordered) response vector with p categories, where the observed category has a value 1 and the remainder are zero. For each category h, $h = 1, \ldots, p$ we assume that an underlying latent normal variable y_{hij} exists and that we have the following multivariate model for these:

$$\begin{aligned} y_{hij} &= X_{ij}\beta_h + z_{hij}u_{hj} + e_{hij} \\ e_{hij} &\sim MVN(0, \Omega_1),\ u_{hj} \sim MVN(0, \Omega_2) \end{aligned} \tag{7.3}$$

The maximum indicant model states that we observe category h for individual ij if $y_{hij} > y_{h^*ij}\ \forall h^* \neq h$. Since an observation must be in one category, one of these latent variables is redundant, say the final one, y_{pij}. This leads to the following conditional sampling scheme (Goldstein *et al*., 2009).

Using the same notation as before, we select a sample of $p-1$ values from $N(X_{ij}\beta_1 + y^*_{ij}\beta_2 + z_{ij}u_j, I_{p-1} - \Sigma_{21}\Sigma_1^{-1}\Sigma_{12})$ and accept this draw to replace the current set of $p-1$ values if and only if the maximum of these $p-1$ values actually occurs in the category where a response variable value of 1 is observed and if this maximum is greater than zero, or if the maximum is less than or equal to zero and a value of 1 is observed in the final category. If not, we select another sample. Note that the sub-matrix corresponding to these $p-1$ latent variables is the identity matrix I_{p-1}.

7.5 Sampling count data

Consider the Poisson distribution with mean θ

$$f(h;\theta) = e^{-\theta}\theta^h/h! \quad h = 0, \ldots p-1 \tag{7.4}$$

for the first p categories that are observed with the cumulative distribution

$$F(h;\theta) = \sum_{g=0}^{h} f(g;\theta)$$

We choose the reference value $h = 0$ and our sampling parallels that for an ordered categorical variable.

For an observed count in the first category we sample a value from the following interval of the normal distribution with variance $(1 - \Sigma_{21}\Sigma_1^{-1}\Sigma_{12})$

$$(-\infty, -(X_{ij}\beta_1 + y^*_{ij}\beta_2 + z_{ij}u_j))$$

and for the remaining categories if we observe a value in category h^* we sample from

$$(a, b), \quad a = \Phi^{-1}[F(h^*; \theta_i)], \quad b = \Phi^{-1}[F(h^* - 1; \theta_i)]$$

where Φ is the cumulative distribution function of $N(0, 1 - \Sigma_{21}\Sigma_1^{-1}\Sigma_{12})$.

Latent normal samplers for other discrete distributions that depend on one or more parameters can be derived in a similar fashion.

7.6 Sampling continuous non-normal data

For a wide class of distributions, we can apply a normalising transformation that is a function of one or more parameters, and then incorporate this in a similar way to that described for discrete responses. For example, the Box-Cox transformation (Box and Cox, 1964) for $y \geq 0$ is

$$y^{(\lambda)} = \begin{cases} (y^\lambda - 1)\lambda^{-1}, & \lambda \neq 0 \\ \log(y), & \lambda = 0 \end{cases}$$

and requires a step for sampling the parameter λ and using this to transform the responses. This step can be carried out using MH with a suitable proposal. The following is an outline of the procedure.

The relevant component of the likelihood for the untransformed y is

$$-\sum_i \frac{(y_i^{(\lambda)} - \mu_i)^2}{2\sigma^2} + (\lambda - 1)\sum_i \log(y_i)$$

where μ_i comprises the fixed predictor, including level 2 random effects for level 1 responses, plus conditioning on the remaining random effects at the same level and σ^2 is the conditional variance on the transformed scale. The second term in this expression is derived from the Jacobian of the transformation. Starting values for the parameter and a Gaussian proposal distribution variance can be obtained from an initial maximum likelihood estimation for the relevant variable.

Where it is not possible to find a suitable transformation of this kind a practical procedure is to divide the measurement scale for the variable into groups of approximately equal size and treat the grouping as an ordered categorical response. While this loses some information due to grouping, if there is a large enough number of groups and a smoothing function can be used, then this may be practically feasible.

7.7 Sampling the level 1 and level 2 covariance matrices

Where all the responses are normal, we can sample the level 1 and level 2 covariance matrices using an inverse Wishart distribution as described in Appendix 6.1. Where

we have categorical responses at level 1 the covariance matrix contains some elements fixed at 0 or 1 and we cannot use an inverse Wishart sampler. For all the categorical responses, the level 1 variances are fixed to be equal to 1.0, with zero correlations among the categories of each unordered categorical variable, but nonzero correlations between these categories and other categorical and continuous variables. Thus, for this set of correlations and for the unconstrained variances, we use an MH sampling procedure.

We assume uniform priors.

Let $\Omega_{1,lm}$ denote the *l,m*-th element of the covariance matrix. We update these covariance parameters using a Metropolis step and a Gaussian random walk proposal, as follows.

At iteration t generate $\Omega^*_{1,lm} \sim \mathrm{N}(\Omega^{(t-1)}_{1,lm}, \sigma^2_{plm})$ where σ^2_{plm} is a proposal distribution variance that has to be set for each covariance and variance. Then form a proposed new matrix Ω^*_1 by replacing the *l,m* th element of $\Omega^{(t-1)}_1$ by this proposed value unless Ω^*_1 is not positive definite in which case set $\Omega^{(t)}_{1,lm} = \Omega^{(t-1)}_{1,lm}$. That is set $\Omega^{(t)}_{1,lm} = \Omega^*_{1,lm}$ with probability $\min[1, p(\Omega^*_1|e_{ij})/p(\Omega^{(t-1)}_1|e_{ij})]$ and $\Omega^{(t)}_{1,lm} = \Omega^{(t-1)}_{1,lm}$ else. The components of the acceptance ratio are

$$p(\Omega^*_1|e_{ij}) = \prod_{ij} |\Omega^*_1|^{-1/2} \exp(-(e_{ij})^T(\Omega^*_1)^{-1}e_{ij}/2)$$

and

$$p(\Omega^{(t-1)}_1|e_{ij}) = \prod_{ij} |\Omega^{(t-1)}_1|^{-1/2} \exp(-(e_{ij})^T(\Omega^{(t-1)}_1)^{-1}e_{ij}/2)$$

We can use an adaptive procedure (Browne *et al*., 2002) to select the proposal distribution parameters.

7.8 Model fi

We may use of the Deviance Information Criterion (DIC) (Spiegelhalter *et al*., 2002) discussed in Chapter 2 to assess model fit. This requires the calculation of the likelihood at each cycle of the MCMC algorithm. To do this we can write the full set of responses as

$$Y = \begin{Bmatrix} Y_1 \\ Y_2 \\ Y_3 \end{Bmatrix}$$

where Y_1, Y_2, Y_3 refer respectively to the normal, ordered and unordered categorical sets of variables. We write the likelihood as

$$P(Y) = P(Y_1|Y_2, Y_3)P(Y_2|Y_3)P(Y_3)$$

The first term on the right-hand side is the multivariate normal likelihood adjusted for the remaining responses. Where we have a non-normal continuous response we

can include this, using the expression given in Section 7.6. The computations for the second term can be carried out by writing $P(Y_2|Y_3)$ as a product of individual conditional variables so that we only require the evaluation of univariate density intervals. Evaluation of the third term involves evaluating the joint distribution of relevant order statistics, and this again can be expressed as a set of conditional distributions, one for each categorical variable.

A practical procedure for carrying out these computations is simulating from the model defined by the current parameter estimates at each iteration to approximate the likelihood.

7.9 Partially ordered data

In the ordered response case where, as described above, we cannot assume underlying multivariate normality, we could treat all categories as unordered but this will lose the information about the ordering. Instead, we can seek to retain as much of this as possible by fitting a partial ordering as follows.

Suppose we have a p-category ordered response. Preserving the ordering, we can split this into two super-categories in $p-1$ ways. We can, in general, split into more than two such categories, but we restrict ourselves here to the simpler case. Suppose we split at category boundary h so that supercategory 1 (S_1) contains the first h categories. We now fit this first set as h unordered categories and the last $p-h$ categories (S_2) as a single further unordered category within which the categories are ordered. Thus in S_2 we fit $p-h-1$ threshold parameters to describe the ordering within this set. Denote these by $\alpha^{S_2} = \left\{\alpha_1^{S_2}, \ldots \alpha_{p-h-1}^{S_2}\right\}$, $\alpha_1^{S_2} = 0$. In practice, we may wish to try several choices for h. We note that we can exchange the two sets of categories if we wish and corresponding procedures apply for splits into more than two sets. While we have introduced this as a device for improving the fit to a latent multivariate normal distribution, there are data types that naturally conform to such a structure. One such example is where we have a set of disease categories where one (or more) can also be further graded by severity.

A further step is introduced into the estimation algorithm. The $h+1$ categories consisting of the first h unordered categories and S_2 are treated as in the completely unordered case with, for convenience, S_2 chosen as the final category. If the observed category belongs to S_1, we sample from the corresponding normal distribution *if* the values sampled from the $h+1$ variate normal distribution values include at least one positive value and the maximum corresponds to the category observed. If all the values sampled are negative, we sample according to the observed S_2 category using the current values of α^{S_2} as in the ordered category case. The α^{S_2} are then sampled in a further step.

7.10 Hybrid normal/ordered variables

In some cases, a variable has observed values that can be expressed either as continuous (normal) or on an ordered scale where there is an underlying normal distribution

that has the same parameters as the observed normal variables. Such data may arise where there are different instruments carrying out essentially the same measurements, or, for example, where different observers choose to measure on either a continuous or on an ordered scale. We here assume that the 'choice' of which values are continuous and which ordered is random, at least conditionally. We assume also that we have some values measured on the continuous scale. We distinguish two situations.

7.10.1 Ordered data with known thresholds

In this situation, the normal scale is grouped and the ordered data are reported by group. The cut points (α_i) therefore are known (or at least to a good approximation). Suppose we have p groups numbered 0,1,...$p-1$. Where we have an observation in group k, we sample the latent normal variable y from the posterior distribution defined by sampling from $N(\mu, \sigma_y^2)$ in the range $[\alpha_{k-1}, \alpha_k]$, where the mean μ is given by the current fixed part + random effects in the model and the variance is the current variance of y. Note that information about the latter is obtained from both the observed normal values and the sampled values y. If we have several responses, the sampling of y is conditional as described above.

A convenient way to implement the sampling of y is as follows:

- Assume that we have an observation in interval k. Treat the observation as missing (at random) and sample from the posterior distribution, as described in Appendix 6.1.
- If the sampled value lies in interval k, accept it; if not, draw another sample and repeat until we obtain one in interval k. This can also be formulated as placing a value 1 in interval k and zero elsewhere, and then accepting each proposed sampled value with probability equal to these values. A generalisation of this procedure for 'partially known values' is set out in Chapter 16.
- The formulation is quite adaptable. Thus, for example, we may have an observation that can only be classified into a set of adjacent intervals – this may arise, for instance, when different degrees of digit preference occur or different instruments are used. Then we simply give each such sub-interval an equal probability with the probabilities summing to 1.

A problem with this approach is that there are cases where we do not know the cut points, or where the cut points may vary from record to record, for example, according to covariate values, in which case we will need to estimate the cut points in the following way: we retain the above MCMC steps but introduce a further step to estimate the (α_i) using the same MH step as described for ordered data.

We note that for such hybrid data, compared to the standard ordered case, the latent normal distribution is no longer $N(0,1)$ but $N(\mu, \sigma_y^2)$ where the parameters are updated at each cycle.

7.11 Discussion

We have set out a general procedure for joint modelling of mixed response types by formulating transformations to an underlying multivariate normal distribution. The procedure has similarities with the Gaussian copula model (see, for example, Pitt *et al*., 2006). In this model, a set of p latent variables following a $(p \times p)$ multivariate normal distribution have a one-to-one mapping onto p marginal discrete or continuous distributions which are modelled as functions of covariates. The latent normal procedure can be viewed as a multilevel generalisation of this copula model, and it also extends the standard copula model to allow the incorporation of multinomial marginal distributions.

We show in Chapter 16 how the model can be extended to deal with responses that occur at more than one level or classification and we also show how the procedure can be used to carry out very general multiple imputations.

8

Multilevel factor analysis, structural equation and mixture models

8.1 A 2-stage 2-level factor model

The theory and application of single level structural equation models (SEMs), including the special cases of observed variable path models and factor analysis models, is well known (Joreskog and Sorbom, 1979, McDonald, 1985). In considering multilevel generalisations of these models, we start with a factor analysis model, which is then elaborated. Early work on estimation procedures based upon maximum likelihood are set out in Goldstein and McDonald (1987), McDonald and Goldstein (1988) with elaborations by Muthen (1989) and Longford and Muthen (1992). Raudenbush (1995) applied the EM algorithm to estimation for a 2-level structural equation model and Rowe and Hill (1997) show how standard multilevel software can be used to provide approximations to maximum likelihood estimates in general multilevel structural equation models. A detailed, likelihood based, treatment of multilevel SEMs is given by Skrondal and Rabe-Hesketh (2004).

We shall now describe general approaches based both upon maximum likelihood and MCMC methods.

Consider first a basic 2-level factor model where we have, say, a set of measurements on each student within a sample of schools together with a set of measurements at the school level which may include aggregated student level measurements. The response measurements of interest whose structure we wish to explore are assumed to be (multivariate) normally distributed random variables. A further set of covariates, for example, gender or social class, are explanatory variables which we may wish

Multilevel Statistical Models: 4th Edition Harvey Goldstein

to condition on. For the p level 1 responses we first write a multivariate model with p responses, where in general some may be missing at random:

$$y_{rij} = (X\beta)_{rij} + z_{rij}e_{rij} + z_{rj}u_{rj} \tag{8.1}$$

where r indexes the response. This can fitted as a 3-level model with the different responses defining level 1, as described in Chapter 6, with dummy variables z_{rij}, z_{rj} for each response and with random coefficients e_{rij}, u_{rj} at levels 2 and 3. Note that at level 3 (between-schools), some of the responses may not vary. Note also that in general some of the coefficients of the covariates may vary at level 3 and would be incorporated as further level 3 random variables along with those above. In terms of the original 2-level model, we now have a set of level 1 random variables e_{rij} and a set of level 2 random variables u_{rj}. A general factor structure for the level 1 variables may involve factors defined at both level 1 and level 2, where we can write

$$\begin{aligned} e_{rij} &= \textstyle\sum_g \lambda_{rg}^{(1)} f_{gij}^{(1)} + w_{rij}^{(1)} \\ u_{rj} &= \textstyle\sum_g \lambda_{rg}^{(2)} f_{gj}^{(2)} + w_{rj}^{(2)} \end{aligned} \tag{8.2}$$

where, for simplicity, we have assumed g factors at each level, although in general we can have different numbers of factors at each level. The λ, f, w respectively refer to the loadings, factors and residuals or 'uniquenesses', and the superscripts indicate the factor levels. In some cases, we may wish to identify some of these factors as the 'same' factors at each level, for example, by constraining certain loadings to be equal. Thus, we may have an attitude score with no between-school variation and any aggregate level variables, by definition, will not vary between pupils within schools. The latter, nevertheless, may enter the model with the level 1 random variables as responses, by being part of the level 2 factor structure and contributing to the prediction of the u_{rj} in the above equation. We can in principle consider any level 2 random variables including random coefficients of covariates when modelling the factor structure at this level.

A straightforward and consistent procedure for estimating the parameters of this factor model is to do it in two stages. The first stage involves the estimation of the separate level 1 and level 2 residual covariance matrices for the responses, using the procedures given in Chapter 6, including cases where there are randomly missing values for some of the responses. The second stage involves the factor analysis of these separate matrices using any standard procedure, as described for example in Joreskog and Sorbom (1979) or McDonald (1985). McDonald (1993) gives details for maximum likelihood estimators in this case.

The 2-stage procedure should be reasonably efficient except where the data are unbalanced, with highly variable numbers of level 1 units within level 2 units. It has the advantage that it can be used for quite general structures. Thus it extends straightforwardly to any number of hierarchical levels. Furthermore, we can also fit models where there are random cross classifications using the procedures described in Chapter 12. Thus, if students are classified by the primary and secondary

school they attended, we can estimate the covariance matrices for level 1 and for both classifications at level 2 and then carry out three separate factor analyses of these matrices. Rowe and Hill (1997) describe this procedure applied to some educational data.

This procedure also allows us to fit general unconditional path models, with or without latent variables, since the covariance matrices at each level are sufficient for these models. A simple example of such a model for a bivariate response without latent variables is as follows

$$\begin{aligned} y_{1ij} &= \alpha_1 + \beta_1 x_{1ij} + u_{1j} + e_{1ij} \\ y_{2ij} &= \alpha_2 + \beta_2 y_{1ij} + u_{2j} + e_{2ij} \end{aligned} \tag{8.3}$$

where the y_{1ij} appears in both equations. The traditional path model treats y_{1ij} in the second of these equations conditionally, so that it can be treated straightforwardly as a bivariate 2-level model. A choice between such a model and the unconditional model will depend on substantive considerations, especially where there is a temporal ordering of variables, when the conditional model would seem to be more appropriate in general. McDonald (1985) gives an account of estimation for unconditional path models. We provide a more general specification for such models below.

8.2 A general multilevel factor model

We now look at general models and efficient estimation methods. We can write a general factor model with normal responses as follows:

$$\begin{aligned} &y_{rij} = \beta_r + \sum_k \alpha_{rk} x_{kij} + \sum_{h=1}^{H} \lambda^{(2)}_{rh} f^{(2)}_{hj} + \sum_{g=1}^{G} \lambda^{(1)}_{rg} f^{(1)}_{gij} + u_{rj} + e_{rij} \\ &u_{rj} \sim N(0, \sigma^2_{ur}), \quad e_{rij} \sim N(0, \sigma^2_{er}), \quad f^{(2)}_{j} \sim MVN_H(0, \Omega_2), \\ &f^{(1)}_{ij} \sim MVN_G(0, \Omega_1) \\ &\quad r = 1, \ldots, R, \quad i = 1, \ldots, n_j, \quad j = 1, \ldots, J, \quad \sum_{j=1}^{J} n_j = N \end{aligned} \tag{8.4}$$

Here we have R responses for N individuals split between J level 2 units. We have H sets of factors, $f^{(2)}_{hj}$ defined at level 2 and G sets of factors, $f^{(1)}_{gij}$ defined at level 1. In the fixed part of the model, we fit separate intercept terms β_r for each response and allow covariates x_{kij}. The loadings, $\lambda^{(1)}$, $\lambda^{(2)}$ are specific to each level. It is possible to further elaborate this model by considering a factor structure for any random coefficients α_{rkj} but we will not pursue this. The residuals, or 'uniquenesses' at levels 1 and 2, e_{rij} and u_{rj} are assumed to be mutually independent, that is, having diagonal covariance matrices.

Skrondal and Hesketh (2004) discuss this model and extensions to structural equation models (see below) as well as models where the responses are discrete or mixtures of discrete and continuous responses with various link functions. They provide maximum likelihood procedures using quadrature estimation, which have

been implemented in the GLLAMM software (Rabe-Hesketh and Skrondal, 2005). (See also Chapter 18.)

Although (8.4) allows a very flexible set of factor models, it should be noted that in order for such models to be identifiable suitable constraints must be placed on the parameters. These will generally consist of fixing the values of some of the elements of the factor variance matrices, Ω_1 and Ω_2 and/or some of the factor loadings, $\lambda_{fr}^{(2)}$ and $\lambda_{gr}^{(1)}$. See Everitt (1984) for an introductory discussion of identifiability.

If we knew the values of the loadings λ then we could fit (8.4) directly as a 3-level model with the loading vectors as the explanatory variables for level 2 and level 3 random coefficients. Conversely, if we knew the values of the random effects f we could estimate the loadings as fixed coefficients in a multivariate response model. These considerations suggest that an EM algorithm can be used in the estimation where the random effects are regarded as missing data (see Rubin and Thayer, 1982). It also suggests an MCMC approach, which also has the advantage of providing exact inferences for parameters, is readily extensible and can incorporate prior information about parameters.

8.3 MCMC estimation for the factor model

To illustrate MCMC estimation for the factor model, consider the following very simple 1-level factor model first.

$$\begin{aligned} y_{ri} &= \beta_{r0} + \lambda_r f_i + e_{ri} \\ f_i &\sim N(0, 1), \quad e_{ri} \sim N(0, \sigma_{er}^2) \end{aligned} \tag{8.5}$$

We will assume that the factor loadings have Normal prior distributions, $p(\lambda_r) \sim N(\lambda_r^*, \sigma_{\lambda r}^2)$ and that the level 1 variance parameters have independent inverse Gamma priors, $p(\sigma_{er}^2) \sim \Gamma^{-1}(a_{er}^*, b_{er}^*)$ and that the intercepts have uniform priors. We have constrained the variance to 1, which identifies the model and fixes the factor scale. The * superscript is used to denote the appropriate parameters of the prior distributions. This model can be updated using a very simple 4-step Gibbs sampling algorithm, as follows (see also Appendix 2.5).

Step 1 Update $\lambda_r (r = 1, \ldots, R)$ from the following distribution: $p(\lambda_r) \sim N(\hat{\lambda}_r, D_r)$ where

$$D_r = \left(\frac{\sum_i v_i^2}{\sigma_{er}^2} + \frac{1}{\sigma_{\lambda r}^2} \right)^{-1}$$

and

$$\hat{\lambda}_r = D_r \left(\frac{\sum_i v_i y_{ri}}{\sigma_{er}^2} + \frac{\lambda_r^*}{\sigma_{\lambda r}^2} \right)$$

Step 2 Update $v_i (i = 1, \ldots, N)$ from the following distribution: $p(v_i) \sim N(\hat{v}_i, D_i)$ where

$$D_i = \left(\frac{\sum_r \lambda_r^2}{\sigma_{er}^2} + 1 \right)^{-1}$$

and

$$\hat{v}_i = D_i \left(\frac{\sum_r \lambda_r^2 y_{ri}}{\sigma_{er}^2} \right)$$

Step 3 Update σ_{er}^2 from the following distribution: $p(\sigma_{er}^2) \sim \Gamma^{-1}(\hat{a}_{er}, \hat{b}_{er})$ where $\hat{a}_{er} = N/2 + a_{er}^*$ and $\hat{b}_{er} = \frac{1}{2} \sum_i e_{ri}^2 + b_{er}^*$.

Step 4 Update β_{r0} from the following distribution: $p(\beta_{r0}) \sim N(\hat{\beta}_{r0}, D_{r\beta})$ where

$$D_{r\beta} = \sigma_{er}^2 / N$$
$$\hat{\beta}_{r0} = \sum_i (y_{ri} - \lambda_r f_i)/N.$$

Goldstein and Browne (2005) show how the MCMC algorithm can be extended for general multilevel models, which allow both orthogonal and correlated factor structures using Gibbs and MH sampling. They also propose a goodness of fit statistic for model comparison. Ansari and Jedidi (2002) also give a general exposition of this model and provide a model comparison procedure.

8.3.1 A 2-level factor example

We use the Science attainment test score data described in Chapter 6 to illustrate the application of two 2-level factor models. The results for two models are given in Table 8.1. We omit the fixed effects estimates since they are very close to those in Table 6.4. Model A has two factors at level 1 and a single factor at level 2. For identification, we have constrained all the variances to be 1.0 and allowed the covariance (correlation) between the level 1 factors to be estimated. Inspection of the correlation structure suggests a model where the first factor at level 1 has nonzero loadings for Earth Science and Biology, constraining those for Physics to be zero (the physics responses have the highest correlation), and for the second factor at level 1 to allow only the loadings for Physics to be unconstrained. The high correlation of 0.90 between the factors suggests that perhaps a single factor will be an adequate summary. Although we do not present results, we have also studied a similar structure for two factors at the school level where the correlation is estimated to be 0.97, strongly suggesting a single factor at that level.

For model B we have separated the three topics of Earth Science, Biology and Physics to separately have nonzero loadings on three corresponding factors at the student level. This time the high inter-correlation is that between the Biology and

Table 8.1 Science attainment 2-level factor model: MCMC estimates.

Parameter	Estimate (s.e.)	Estimate (s.e.)
	Model A	Model B
Level 1: factor 1 loadings		
Earth Science core	0.06 (0.004)	0.11 (0.02)
Biology core	0.11 (0.004)	0*
Biology R3	0.05 (0.008)	0*
Biology R4	0.11 (0.009)	0*
Physics core	0*	0*
Physics R2	0*	0*
Level 1; factor 2 loadings		
Earth Science core	0*	0*
Biology core	0*	0.10 (0.005)
Biology R3	0*	0.05 (0.008)
Biology R4	0*	0.10 (0.009)
Physics core	0.12 (0.005)	0*
Physics R2	0.12 (0.007)	0*
Level 1; factor 3 loadings		
Earth Science core	–	0*
Biology core	–	0*
Biology R3	–	0*
Biology R4	–	0*
Physics core	–	0.12 (0.005)
Physics R2	–	0.12 (0.007)
Level 2; factor 1 loadings		
Earth Science core	0.04 (0.007)	0.04 (0.007)
Biology core	0.09 (0.008)	0.09 (0.008)
Biology R3	0.05 (0.009)	0.05 (0.010)
Biology R4	0.10 (0.016)	0.10 (0.016)
Physics core	0.10 (0.010)	0.10 (0.010)
Physics R2	0.09 (0.011)	0.09 (0.011)
Level 1 residual variances		
Earth Science core	0.017 (0.001)	0.008 (0.004)
Biology core	0.015 (0.001)	0.015 (0.001)
Biology R3	0.046 (0.002)	0.046 (0.002)
Biology R4	0.048 (0.002)	0.048 (0.002)
Physics core	0.016 (0.001)	0.016 (0.001)
Physics R2	0.029 (0.002)	0.030 (0.002)

Table 8.1 (*Continued*)

Parameter	Estimate (s.e.)	Estimate (s.e.)
Level 2 residual variances		
Earth Science core	0.002 (0.0005)	0.002 (0.0005)
Biology core	0.0008 (0.0003)	0.0008 (0.0003)
Biology R3	0.002 (0.0008)	0.002 (0.0008)
Biology R4	0.010 (0.002)	0.010 (0.002)
Physics core	0.002 (0.0005)	0.002 (0.0005)
Physics R2	0.003 (0.0009)	0.003 (0.0009)
Level 1 correlation factors 1 &2	0.90 (0.03)	0.55 (0.10)
Level 1 correlation factors 1 &3	–	0.49 (0.09)
Level 1 correlation factors 2 &3	–	0.92 (0.04)

*indicates constrained parameter. A chain of length 20 000 with a burn in of 2000 was used. Level 1 is student, level 2 is school.

Physics booklets with only moderate (0.49, 0.55) correlations between the factors for Earth Science and each of Biology and Physics. In fact this model is not properly identified because of the single loading for Earth Sciences Core at level 1. The correlation estimates are strongly dependent on the chosen prior, in this case for the correlations the prior is uniform over (–1,1) and a different choice of prior, for example uniform over (0,1) will lead to different estimates. In the extreme case a point prior is equivalent to fixing the value of the correlation to a given value. In general we can use informative, other than point, priors for any of the parameters in the model and this provides an alternative method of handling identifiability to that of fixing parameters at particular values.

The results suggest that we need at least two factors to describe the student level data and that the preliminary analysis using just one factor can be improved upon.

8.4 Structural equation models

In the basic factor model, the factors themselves are not modelled any further. In many applications, however, we may wish to do this and these models are generally referred to as structural equation models. We first look at one simple extension to the 2-level factor model with responses at a single level and a factor structure fitted at level 1 where the factor is a linear function of explanatory variables.

$$\begin{aligned} y_{rij} &= \beta_r + \lambda_r^{(1)} f_{ij}^{(1)} + e_{rij}, \quad f_{ij}^{(1)} = \alpha_1 x_{ij} + u_j + w_{ij} \\ w_{ij} &\sim N(0, \sigma_w^2), \quad e_{rij} \sim N(0, \sigma_e^2), \quad u_j \sim N(0, \sigma_u^2) \end{aligned} \tag{8.6}$$

where, for identifiability, we set $\sigma_w^2 = 1$.

On substitution, this gives

$$y_{rij} = \beta_{0r} + \lambda_r^{(1)}\alpha_1 x_{ij} + \lambda_r^{(1)} u_j + \lambda_r^{(1)} w_{ij} + e_{rij}.$$

The MCMC algorithm now involves an extra step to sample α_1 which is similar to the step for sampling the $\lambda_r^{(1)}$. Details are given in Goldstein *et al*. (2007a). We can fit a level 2 structure and further explanatory variables in either the fixed or structural part of the model, as well as further factors, and this is elaborated below.

We now look at a more general structural equation model that allows relationships among factors. We illustrate using a single level model that can be written in the following matrix form (see McDonald, 1985, for some alternative representations):

$$\begin{aligned} A_1 f_1 &= A_2 f_2 + W \\ Y_1 &= \Lambda_1 f_1 + U_1 \\ Y_2 &= \Lambda_2 f_2 + U_2 \end{aligned} \tag{8.7}$$

Where Y_1, Y_2 are observed multivariate vectors of responses, A_1 is a known transformation matrix, often set to the identity matrix, A_2 is a coefficient matrix which specifies a multivariate linear model between the set of transformed factors, f_1, f_2. Λ_1, Λ_2 are loadings, U_1, U_2 are uniquenesses, W is a random residual vector and W, U_1, U_2 are mutually independent, with zero means. The extension of this model to the multilevel case follows that of the factor model; we shall restrict ourselves to sketching how an MCMC algorithm can be applied to (8.7). Note that we can write A_2 as the vector A_2^* by stacking the rows of A_2. For example, if

$$A_2 = \begin{pmatrix} a_0 & a_1 \\ a_2 & a_3 \end{pmatrix}, \text{ then } A_2^* = \begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{pmatrix}$$

The distributional form of the model can be written as

$$\begin{aligned} &A_1 f_1 \sim MVN(A_2 f_2, \Sigma_3) \\ &f_1 \sim MVN(0, \Sigma_{f_1}), \;\; f_2 \sim MVN(0, \Sigma_{f_2}) \\ &Y_1 \sim MVN(\Lambda_1 f_1, \Sigma_1), \quad Y_2 \sim MVN(\Lambda_2 f_2, \Sigma_2) \end{aligned}$$

with priors

$$A_2^* \sim MVN(\hat{A}_2^*, \Sigma_{A_2^*}), \quad \Lambda_1 \sim MVN(\hat{\Lambda}_1, \Sigma_{\Lambda_1}), \quad \Lambda_2 \sim MVN(\hat{\Lambda}_2, \Sigma_{\Lambda_2})$$

and Σ_1, Σ_2, Σ_3 having inverse Wishart priors.

The coefficient and loading matrices have conditional normal distributions, as do the factor values. The covariance matrices and uniqueness variance matrices involve steps similar to those given in the earlier algorithm. The extension to two and more levels follows readily.

The model can be generalised further by considering m sets of response variables, $Y_1, Y_2, \ldots Y_m$ in (8.7) and several, linked, multiple group structural relationships with the k-th relationship having the general form

$$\sum_h V_h^{(k)} A_h^{(k)} = \sum_g V_g^{(k)} A_g^{(k)} + W^{(k)}$$

and the above procedure can be extended for this case. Note that the model for simultaneous factor analysis (or, more generally, structural equation modelling) in several populations is a special case of this model, with the addition of any required constraints on parameter values across populations.

As suggested above, we can elaborate our models to include fixed effects, responses at level 2 as well as level 1, and covariates Z_1, Z_2 for the factors, which may be a subset of the fixed effects covariates X, for example,

$$\begin{aligned} Y^{(1)} &= X\beta + \Lambda_2^{(1)} f_2 Z_2^{(1)} + u^{(1)} + \Lambda_1^{(1)} f_1 Z_1^{(1)} + e^{(1)} \\ Y^{(2)} &= \Lambda_2^{(2)} f_2 Z_2^{(2)} + u^{(2)} \\ Y^{(1)} &= \{y_{rij}\}, \quad Y^{(2)} = \{y_{rj}\} \\ r &= 1, \ldots, R, \quad j = 1, \ldots, J \end{aligned} \tag{8.8}$$

The superscript refers to the level at which the measurement exists, so that, for example, y_{1ij}, y_{2j} refer respectively to the first measurement in the i-th level 1 unit in the j-th level 2 unit (say, students and schools) and the second measurement taken at school level for the j-th school.

8.5 Discrete response multilevel structural equation models

Suppose we have a set of binary responses for a set of variables measured on individuals. We can write a model analogous to (8.5) as

$$\begin{aligned} &f(\pi_{ri}) = \lambda_r \nu_i + e_{ri}, \ r = 1, \ldots, R, \ i = 1, \ldots, N \\ &\nu_i \sim N(0, 1), \ e_{ri} \sim N(0, \sigma_{er}^2) \\ &y_{ri} \overset{iid}{\sim} bin(1, \pi_{ri}) \end{aligned} \tag{8.9}$$

where $\pi_{ri} = \Pr(y_{ri} = 1)$, and for convenience we choose a probit link function so that the unique variances, σ_{er}^2, are fixed at 1. As before, we can use a Gibbs sampler (see Appendix 4.3) where the lowest level variables are also thought of as latent, and generalisations to several levels and to structural equations follow similar lines to those described above. Software (for example, REALCOM, WINBUGS and GLLAMM) is available to fit such models, including those with mixtures of normal and ordered responses.

Model (8.9) is often referred to as an item response model (IRM) (see Goldstein and Wood, 1989) and (8.9) and its extensions allow a very flexible class of models to be fitted. Thus, for an ordered response model, for example, sometimes termed a 'partial credit' model, we can also use the probit link and can therefore fit mixtures of binary, ordered and continuous responses in the same model. Likewise, Poisson count responses can be incorporated using the latent normal formulation for this given in Chapter 7. Goldstein *et al*. (2007a) fit a 2-level binary response factor model with a probit link to a set of 15 Mathematics items from a test given to French and British school students. They show how country differences can be modelled and how multiple factors, for each country, at both levels can be fitted. They provide estimation details and a DIC goodness of fit test.

8.6 More complex hierarchical latent variable models

In Chapter 6, we saw how we could simultaneously model responses at several levels. We can extend the factor model and the structural equation model in similar ways. Consider, for example, the structural equation model defined by (8.6), where we have an additional set of level 2 responses,

$$\begin{aligned} &y_{rij} = \beta_r^{(1)} + \lambda_r^{(1)} v_{ij}^{(1)} + e_{rij}, \quad v_{ij}^{(1)} = \alpha_1 x_{ij} + u_j^{(1)} + w_{ij} \\ &y_{rj} = \beta_r^{(2)} + \lambda_r^{(2)} v_j^{(2)} + u_{rj}^{(2)} \\ &w_{ij} \sim N(0, \sigma_w^2), \quad e_{rij} \sim N(0, \sigma_e^2) \\ &\begin{pmatrix} u_j^{(1)} \\ u_{rj}^{(2)} \end{pmatrix} \sim N(0, \Omega_u) \end{aligned} \tag{8.10}$$

We can introduce further classifications, either hierarchical or crossed, each with their own structure.

In these latent variable models, as we introduce explanatory variables, so the interpretation of the factors will change since we are adjusting for these variables. We can consider fitting a 'standardising model', for example, with no explanatory variables and subsequently treating the estimates of the loadings from this model as fixed in further analyses. The advantage of this approach is that we are dealing with essentially the same factors, as defined by the loadings, in each analysis. The details for imposing constraints are given in Goldstein *et al*. (2007a).

8.7 Multilevel mixture models

In some applications it may be reasonable to consider that units at a particular level of the data hierarchy actually come from more than one population and that the model parameters vary across populations. We saw one such example, the mover-stayer model in Section 5.10, in which an individual was in one of three population

groups and interest centres on estimating the parameters associated with each group and the proportions in each group. A more general case is where the number of groups is unknown and where we may also have a model for the probability of group membership that is dependent on explanatory variables. In the 2-level case, we assume that each level 1 unit may belong to a class ($c = \{c_k\}$) defined at level 2 and for each class we have a different relationship, as follows:

$$\begin{aligned} y_{ij}^{c_k} &= (X\beta)_{ij}^{c_k} + u_j^{c_k} + e_{ij}^{c_k} \\ u_j^{c_k} &\sim N(0, \sigma_{u,c_k}^2),\ e_{ij}^{c_k} \sim N(0, \sigma_{e,c_k}^2) \end{aligned} \tag{8.11}$$

The class allocation model can be written as a function of level 2 variables, for example,

$$\begin{aligned} pr(ij \in c_k) &= f(\alpha_{k0} + \alpha_{k1} z_j + v_{kj}) \\ v_{kj} &\sim N(0, \sigma_v^2) \end{aligned} \tag{8.12}$$

Within each class we thus have different random and fixed parameters; the probability of class membership is also a function of fixed and random effects. The model can be extended by introducing random coefficients and further explanatory variables, and also by allowing classes defined at level 1 (see Vermunt, 2008). Such models have often been applied to repeated measures data, where each level 2 unit is considered to belong to a group with its own growth trajectory (see for example Nagin, 1999, Leiby *et al*., 2009). Muthen and Asparouhov (2009) discuss 2-level mixture models with various examples. If maximum likelihood estimation is used, then the number of groups may be chosen by fitting models with different numbers of groups and using the AIC or BIC criterion for choosing between them (Section 2.12.3). We note, however, as with generalised linear models, maximum likelihood estimation can be very slow if there are large numbers of random parameters and the occurrence of multiple local maxima may also be a problem. We may also allow $u_j^{c_k}, v_j$ to be correlated and this can be useful if the class allocation model is not fully specified.

It is often possible to fit a given dataset equally well using a model such as (8.11), (8.12) and a standard 2-level model with random coefficients or a model with complex level 1 variance. These models will have different interpretations so that some care needs to be exercised in interpreting these models.

MCMC estimation can be carried out for such models and the steps, in outline for (8.11) and (8.12), are as follows for a fixed number of groups. Default diffuse priors can be used, but we will often have information about the proportions in each group that can be used as informative priors.

At each cycle each level 1 unit is allocated to a class and for each such class the parameters are sampled in the usual way for a 2-level model. If there are fixed or random parameters in common across classes, then for these parameters the relevant classes are combined. This may lead to either the fixed or random parameters needing to be updated one at a time.

The parameters of (8.12) are updated with a MH step where the likelihood is given by

$$\prod_{ij} L(y_{ij}^{c_k}) pr(ij \in c_k) \tag{8.13}$$

and in a final step, each level 1 unit in turn is assigned to the group for which (8.13) is maximised. One of the problems that can be encountered is that of label switching where the parameters associated with one particular group become switched with another group during the course of the cycles. This may avoided if it is possible for a few level 1 units to be fixed as members of each group. Lenk and DeSarbo (2000) discuss a general procedure using a particular prior distribution specification, to avoid this problem. If the model is run with different numbers of groups we may use the DIC to decide on the most suitable model.

9

Nonlinear multilevel models

9.1 Introduction

The models of Chapters 2 and 3 are linear in the sense that the response is a linear function of the parameters in the fixed part and the elements of the full covariance matrix V are linear functions of the parameters in the random part. In many applications, however, it is appropriate to consider models where the fixed or random parts of the model, or both, contain nonlinear functions. For example, in the study of growth, Jenss and Bayley (1937) proposed the following function to describe the growth in height of young children

$$\begin{aligned} y_{ij} = \alpha_0 + \alpha_1 t_{ij} + u_{\alpha 0j} + u_{\alpha 1j} t_{ij} + e_{\alpha ij} \\ - \exp(\beta_0 + \beta_1 t_{ij} + u_{\beta 0j} + u_{\beta 1j} t_{ij} + e_{\beta ij}) \end{aligned} \tag{9.1}$$

where t_{ij} is the age of the j-th child at the i-th measurement occasion. Models for discrete data, such as counts or proportions are a special case of nonlinear models often termed 'generalised linear models' (discussed these in Chapter 4). For example, a 2-level log linear model can be written

$$E(m_{ij}) = \pi_{ij}, \quad \pi_{ij} = \exp(X_{ij}\beta_j) \tag{9.2}$$

where m_{ij} is assumed typically to have a Poisson distribution, in this case across level 1 units. In this chapter we consider nonlinear models more generally.

9.2 Nonlinear functions of linear components

The following results are an extension of those presented by Goldstein (1991) (Appendix 9.1 gives details). We will give procedures for deriving likelihood based

Multilevel Statistical Models: 4th Edition Harvey Goldstein

estimates and note that where the random effects are not part of the nonlinear function, the procedure gives maximum likelihood estimates; otherwise these are quasilikelihood estimates similar to those considered in Chapter 4.

MCMC methods can also be used to fit these models. As with generalised linear models (see Appendix 4.3), MH sampling can be used, since not all the distributions can be specified analytically. Otherwise, the steps are as in Appendix 2.5. For some examples, see Gilks *et al*. (1996). The likelihood based estimates can be used to provide starting values for an MCMC estimation.

Restricting attention to a 2-level structure we can write a fairly general model, as follows,

$$y_{ij} = X_{1ij}\beta_1 + Z^{(2)}_{1ij}u_{1j} + Z^{(1)}_{1ij}e_{1ij} + f(X_{2ij}\beta_2 + Z^{(2)}_{2ij}u_{2j} + Z^{(1)}_{2ij}e_{2ij}) + \ldots \qquad (9.3)$$

where the function f is nonlinear and where the $+\ldots$ indicates that additional nonlinear functions can be included, involving further fixed part explanatory variables X or random part explanatory variables at levels 1 and 2, respectively $Z^{(1)}, Z^{(2)}$. The model is first linearised by a suitable Taylor series expansion and this leads to consideration of a linear model where the explanatory variables in f are transformed using first and second derivatives of the nonlinear function. Note that the linear component of (9.3) is treated in the standard way, and that the random variables at a given level in the linear and nonlinear components may be correlated.

Consider the nonlinear function f. Appendix 9.1 gives details to show that we can write this as the sum of a fixed part component and a random part using a Taylor series approximation. Concentrating on the fixed part, this gives

$$f_{ij}(H_{t+1}) = f_{ij}(H_t) + X_{2ij}(\beta_{2,t+1} - \beta_{2,t})f'_{ij}(H_t) \qquad (9.4)$$

where $\beta_{1,t+1}, \beta_{1,t}$ are the current and previous iteration values of the fixed part coefficients and H_t represents the fixed part predictor in the nonlinear component of (9.3). We can choose H_t to be either the current value of the fixed part predictor, that is, $X_{2ij}\beta_2$; or we can add the current estimated residuals to obtain an improved approximation to the nonlinear component for each unit. As with generalised linear models, the former is referred to as a 'marginal' (quasilikelihood) model and the latter as a 'penalised' or 'predictive' (quasilikelihood) model. We note that this procedure reduces to the standard Newton-Raphson scoring algorithm for a single level model (McCullagh and Nelder, 1989).

In practice, general models such as (9.1) may pose considerable estimation problems. We notice that the same explanatory variables can occur in the linear and nonlinear components and this can lead to instability and failure to converge.

9.3 Estimating population means

Consider the expected value of the response for a given set of covariate values. Because of the nonlinearity this is not in general equal to the predicted value when

the random variables in the nonlinear function are zero. For example, if we write the variance components version of (9.2)

$$\pi_{ij} = \exp(\beta_0 + \beta_1 x_{ij} + u_j)$$

and assuming $u_j \sim N(0, \sigma_u^2)$ we obtain the marginal expectation

$$E(\pi_{ij}|x_{ij}) = \exp(\beta_0 + \beta_1 x_{ij}) \int_{-\infty}^{\infty} e^{u_j} \phi(u_j) du_j = \exp(\beta_0 + \beta_1 x_{ij} + \sigma_u^2/2) \qquad (9.5)$$

where ϕ is the density function of the normal distribution. The population predicted values, conditional on covariates, can be obtained if required, as above, by taking expectations over the population. An approximation to this can be obtained from the second order terms in (9.1.4) in Appendix 9.1 with higher order terms introduced if necessary to obtain a better approximation. Alternatively we may generate a large number of simulated sets of values for the random variables and for each set evaluate the response function to obtain an estimate of the full population distribution. See Section 4.9 for more details of such a procedure.

9.4 Nonlinear functions for variances and covariances

We saw in Chapter 3 how we could model complex functions of the level 1 variance. As with the linear component of the model, there are cases where we may wish to model variances or covariances as nonlinear functions. In principle, we can do this at any level but here we restrict our attention to level 1 and to the variance only. In Chapter 5, we gave an example where the covariances are modelled in this way.

Suppose that the level 1 variance decreases with increasing values of an explanatory variable such that it approaches a fixed value asymptotically. We could then model this for a 2-level model, say, as follows:

$$\text{var}(e_{ij}) = \exp(\beta_0^* + \beta_1^* x_{ij})$$

where β_0^*, β_1^* are parameters to be estimated. Such a model also guarantees that the level 1 variance is positive, which is not the case with linear models, such as those based on polynomials. The estimation procedure is analogous to that described above and provides maximum likelihood estimates, and (details are given in Appendix 9.1). In Chapter 17, we extend this to consider models where the variance, and more generally the covariance structure, can be modelled as a function of further variables at any level of a data hierarchy using MCMC estimation; we also consider the case where we treat the level 1 variance as a random effect, varying across level 2 units.

9.5 Examples of nonlinear growth and nonlinear level 1 variance

We give an example of a model with a nonlinear transformation of the linear predictor and then consider the case of a nonlinear level 1 variance function. The example consists of 577 repeated measurements of height on 197 French Canadian boys aged from 5 to 10 years (Demirjian *et al.*, 1982), with between three and seven measurements each; this is a 2-level structure with measurement occasions nested within children.

We fit the following version of the Jenss-Bayley curve to illustrate the procedure

$$y_{ij} = \alpha_0 + \exp(\beta_0 + \beta_1 t_{ij} + \beta_2 t_{ij}^2 + \beta_3 t_{ij}^3 + u_{\beta 0 j} + u_{\beta 1 j} t_{ij}) + e_{\alpha i j} \qquad (9.6)$$

so that the fixed part is an intercept plus a nonlinear component and the random part variance at level 2 is part of the nonlinear component. The results are given in Table 9.1, using the first order approximation with prediction based upon the fixed part only.

The level 1 variance is small and of the order of the measurement error of height measurements. The starting values for this model need to be chosen with care, and in the present case the model was run to convergence without the linear intercept α_0 which was then added with a starting value of 100. Bock (1992) uses an EM algorithm to fit a nonlinear 2-level model to growth data from age 2 years to adulthood using a mixture of three logistic curves.

Our second example is the JSP dataset used in Chapter 3 to study the level 1 variance. We will fit Model B of Table 3.1 with a nonlinear function of the level 1 variance instead of the level 1 variance as a quadratic function of the 8-year-score. This level 1 variance for the ij-th level 1 unit is $\exp(\beta_0^* + \beta_1^* x_{1ij})$ and Table 9.2 shows

Table 9.1 Nonlinear model estimates with first order fixed part prediction. Age is measured about 8.0 years.

Fixed coefficient	Estimate (s.e.)
Intercept (linear)	90.3
Intercept (nonlinear)	3.58
Age	0.15 (0.10)
Age squared	−0.016 (0.02)
Age cubed	0.002 (0.004)

Nonlinear model level 2 covariance matrix (s.e.)

	Intercept	Age
Intercept	0.025 (0.003)	
Age	−0.0027 (0.0003)	0.00036 (0.00005)

Level 1 variance = 0.25

Table 9.2 Nonlinear level 1 variance for JSP data.

Parameter	Estimate (s.e.)
Fixed	
Constant	31.7
8-year score	0.58 (0.03)
Gender (boys-girls)	−0.34 (0.27)
Social class (non-manual-manual)	0.76 (0.30)
School mean 8-year score	0.01 (0.11)
8-year score x school mean 8-year score	0.02 (0.01)
Random	
Level 2	
β_0^*	2.87 (0.88)
β_0^*	−0.17 (0.07)
σ_{u1}^2	0.012 (0.007)
Level 1	
β_0^*	2.74 (0.06)
β_1^*	−0.10 (0.01)

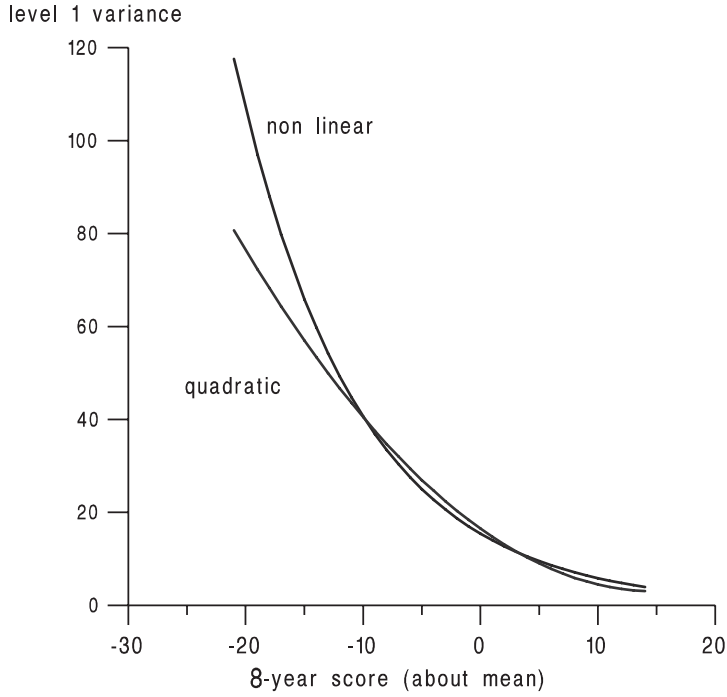


Figure 9.1 Predicted level 1 variance as a function of 8-year Maths score.

the model estimates, which are almost identical to those of Model B of Table 3.1, as is the likelihood value.

Figure 9.1 shows the predicted level 1 variance for this model and Model B of Table 3.1.

In these data, the nonlinear function gives very similar results to the quadratic polynomial one. It is clear, however, that where the variance asymptotically approaches a constant value, for extreme values of an explanatory variable, a linear or even quadratic approximation may be expected to fail. In the present case, a linear function does predict a negative level 1 variance within the range of the data. An example where a nonlinear function is necessary is in growth data, described in Chapter 5, where the level 1 (within-individual) variation will decrease towards a constant value at the approach to adulthood.

Appendix 9.1 Nonlinear model estimation

The following exposition is similar to that in Appendix 4.1 for discrete response (nonlinear) models, but is more general since the level 1 random effects are also part of the nonlinear link function. We consider a single nonlinear term of the form

$$y_{ij} = f(X_{ij}\beta + Z_{2ij}u_{2j} + Z_{1ij}e_{1ij}) \tag{9.1.1}$$

The addition of linear terms to this model is discussed in Chapter 9.

At the $(t+1)$-th iteration in the estimation we expand (9.1.1) for both fixed and random parts as follows

$$\begin{aligned} &f_{ij}(H_t) + X_{ij}(\beta_{t+1} - \beta_t)f'_{ij}(H_t) \\ &+(Z_{2ij}u_{2j} + Z_{1ij}e_{1ij})f'_{ij}(H_t) + (Z_{2ij}u_{2j} + Z_{1ij}e_{1ij})^2 f''_{ij}(H_t)/2 \end{aligned} \tag{9.1.2}$$

in terms of parameter values estimated at the t-th iteration. The first line of (9.1.2) updates the fixed part of the model and in the special case of a single level quasi-likelihood model provides the updating function. The quantity $f_{ij}(H_t) - X_{ij}\beta_t f'_{ij}(H_t)$ is treated as an offset to be subtracted from the response variable. The first term in the second line defines a linear random component based on the explanatory variables transformed by multiplying by the first differential. We need to specify H_t and consider the distribution of the second term in the second line of (9.1.2).

If we choose $H_t = X_{ij}\beta_t$, this is equivalent to carrying out the Taylor expansion around the fixed part predicted value. If we choose $H_t = X_{ij}\beta_t + Z_{2ij}\hat{u}_{2j} + Z_{1ij}\hat{e}_{1ij}$, this expands around the current predicted value for the ij-th unit and we replace the second line of (9.1.2) by

$$\begin{aligned} &(Z_{2ij}(u_{2j} - \hat{u}_{2j}) + Z_{1ij}(e_{1ij} - \hat{e}_{1ij}))f'_{ij}(H_t)+ \\ &(Z_{2ij}(u_{2j} - \hat{u}_{2j}) + Z_{1ij}(e_{1ij} - \hat{e}_{1ij}))^2 f''_{ij}(H_t)/2 \end{aligned}$$

We thus have the further offset from the linear term to be added to the response

$$(Z_{2ij}\hat{u}_{2j} + Z_{1ij}\hat{e}_{1ij})f'_{ij}(H_t)$$

From the second line of (9.1.2) we have

$$\begin{aligned} &E(Z_{2ij}u_{2j} + Z_{1ij}e_{1ij}) = 0, \quad E(Z_{2ij}u_{2j} + Z_{1ij}e_{1ij})^2 = \sigma^2_{zu} + \sigma^2_{ze} \\ &\sigma^2_{zu} = Z_{2ij}\Omega_u Z^T_{2ij}, \sigma^2_{ze} = Z_{1ij}\Omega_e Z^T_{1ij} \end{aligned} \tag{9.1.3}$$

To incorporate the second order terms we treat $(\sigma^2_{zu} + \sigma^2_{ze})f''(H_t)/2$ as an additional offset in the fixed part and in the random part of the model we need to consider the variation of the second term in the second line of (9.1.2). If we assume normality then all third moments, formed from the product of the two terms in the second line

of (9.1.2), are zero and we have

$$\text{var}(Z_{2ij}u_{2ij} + Z_{1ij}e_{1ij})^2 = 2(\sigma_{zu}^4 + \sigma_{ze}^4) \tag{9.1.4}$$

If we define the additional variables

$$Z_u^* = \sigma_{zu}^2 f''(H_t)/\sqrt{2}, Z_e^* = \sigma_{ze}^2 f''(H_t)/\sqrt{2}$$

we form $Z_u^* Z_u^{*T}$, $Z_e^* Z_e^{*T}$ as offsets for the response vector $vec(\tilde{Y}\tilde{Y}^T)$ in the estimation of the random parameters. Having modified the response variable by removing the necessary offsets we are left in the fixed part with a modified response, say Y' with a modified explanatory variable matrix, say X'. We do likewise for the random part of the model and then carry out a standard iterative procedure, updating the differential functions at each iteration.

Where the Taylor expansion is taken about the current values of the fixed part plus residuals we require

$$E(Z_{2ij}(u_{2j} - \hat{u}_{2j}))^2 + E(Z_{1ij}(e_{1ij} - \hat{e}_{1ij}))^2$$

which leads to the 'conditional' or 'comparative' variances described in Appendix 2.2, so that we substitute these variances, $\Omega_{\hat{u}}$ and $\Omega_{\hat{e}}$, for Ω_u and Ω_e in the above expressions for the fixed and random offsets.

To estimate residuals we note that, having adjusted the response using the offsets, we have on the right-hand side of (9.1.2) the fixed part together with the random terms

$$(Z_{2ij}u_{2j} + Z_{1ij}e_{1ij})f'_{ij}(H_t) + ((Z_{2ij}u_{2j} + Z_{1ij}e_{1ij})^2 - (\sigma_{zu}^2 + \sigma_{ze}^2))f''_{ij}(H_t)/2$$

Each residual and its square appear in this expression, and since third order moments are zero, we can apply the usual linear estimation for the residuals as described in Appendix 2.2. The weight matrix V is based upon both the linear and quadratic terms of the above expression. We carry out an analogous procedure for the case where the Taylor expansion is based upon the current fixed part plus residual estimates.

The above can be extended in a straightforward way to more than two levels and to multivariate models.

9.1.1 Modelling variances and covariances as nonlinear functions

In Chapter 2, we saw that the random parameters were estimated by regressing the observed cross-product matrix of residuals on a set of explanatory variables which defined the appropriate variances and covariances at each level. Using the notation in Appendix 2.1 we have the following *linear* model for the random parameters β^*

$$Y^* = vec(\tilde{Y}\tilde{Y}^T) = X^*\beta^*, \quad E(Y^*) = vec(V) \tag{9.1.5}$$

We can now apply the same procedure for the specification and estimation of a nonlinear model as above. We illustrate this for the case where the level 1 variance is an exponential function of a covariate x_1^*, defined in terms of the Kronecker product as in Appendix 2.1, namely for the t-th element of $X^*\beta^*$ (which is on the diagonal of V) we assume a level 1 variance contribution of the form

$$\sigma_{et}^2 = \exp(\beta_0^* + \beta_1^* x_{1t}^*), \beta^* = \begin{pmatrix} \beta_0^* \\ \beta_1^* \end{pmatrix} \tag{9.1.6}$$

As in the linear function case, we form the first differential $f' = f$, multiply x_{0t}^*, x_{1t}^* by this and estimate the parameters of the resulting transformed linear model. This will involve introducing an offset for Y^* and constructing the following level 1 explanatory variables for the estimation of β^*, setting their covariance to zero

$$\{x_{0t}^* \exp(\beta_0^* x_{0t}^* + \beta_1^* x_{1t}^*)\}^{0.5}, \{x_{1t}^* \exp(\beta_0^* x_{0t}^* + \beta_1^* x_{1t}^*)\}^{0.5}$$

Because we are estimating only nonlinear functions of linear components here and not adding approximations to a further random component, the estimates obtained are exact maximum likelihood or restricted maximum likelihood estimates.

In Chapter 5, we developed a special case of a nonlinear model for covariances and in Chapter 9, we gave an example of model (9.1.6). We note that the parameters β_0^*, β_1^* are not necessarily positive when modelling (9.1.6) and although we would normally regard such level 1 parameters as variances, in this case, as in Section 3.1, they are simply parameters to be estimated. As with nonlinear modelling, in general, it is important to have reasonable starting values. These might be obtained by trial and error or by making preliminary estimates of variances for various values of the relevant explanatory variable and regressing their logarithms on the level 1 explanatory variables.

In Chapter 17, we present a more general model using MCMC estimation.

10

Multilevel modelling in sample surveys

10.1 Sample survey structures

Many practical sample surveys have a multistage sample design structure whereby, at the first stage, Primary Sampling Units (PSUs) are chosen randomly from a population of such units. Thus, in a household survey these might be administrative areas. At the next stage, either all the units or a sample of units is drawn from the chosen PSUs and this can lead to a third stage, etc., until the final stage units, the households or persons in this case, are selected. This procedure is often referred to as 'cluster sampling', although strictly this term refers to samples where all the final stage units from each penultimate cluster are chosen. Typically, the selection probabilities at each stage are chosen so that the final sample is *self-weighting*, that is, each final stage unit in the population has the same overall probability of being chosen. For unequal weights, we can use the procedures described in Chapter 3. In Section 10.2.1 we look at how the selection probabilities at each stage can be used in model estimation. In this chapter we shall assume that there are no missing data (see Chapter 16 for procedures to handle such cases).

Sample surveys also often involve stratification, for example, the population might be divided into rural and urban areas and sampling carried out separately for these 'strata', again often so that a self weighting sample is produced. In some cases, stratification may be used to ensure sufficient numbers in a particular category, such as an ethnic minority group. When we come to model a stratified sample we may wish to include in the model some dummy variable terms for the strata categories, and possibly for the categories formed by combinations of strata. Alternatively we may wish to incorporate stratum weights into the overall survey weights for each

Multilevel Statistical Models: 4th Edition Harvey Goldstein

unit. We might wish to do this if we specify a 'marginal' model that describes the population as a whole rather than describing relationships within strata.

An important reason for stratification and clustering is to increase precision for a given cost for a given total sample size. Clustering, for example by area, generally reduces survey costs while increasing standard errors of estimates, while stratification tends to reduce standard errors (see, for example, Kish, 1965), and surveys are typically concerned to achieve a judicious balancing of these two procedures. From a modelling perspective, however, this is of secondary importance since we wish to uncover relationships that exist and estimate covariance structures. Where multistage surveys choose units for administrative and cost-effectiveness reasons, there will usually be little intrinsic interest in the between- and within-unit variation, but it will still usually be efficient to fit a multilevel model. The traditional approach in survey analysis to making inferences about the population, which ignores the existence of any multilevel structure, is to fit models appropriate for single level structures and then to adjust the standard errors to take account of the sample structure (see Kish and Frankel, 1974). Thus, for example, we saw in Section 2.7 that using an OLS estimator when there were cluster differences tended to underestimate the fixed effect standard errors. It is also worth pointing out that, even where a survey involves no clustering or stratification, we may still wish to fit a multilevel model to explore the population data structure.

10.2 Population structures

10.2.1 Superpopulations

A major distinction is between inferences made for finite populations, and those for infinite populations where we make inferences on the basis of standard statistical model assumptions. Much of traditional sample survey theory is concerned with sampling from real, finite populations where we wish to estimate characteristics of the population, such as means or proportions. In a household survey of a large city, we may wish to estimate the actual proportion of one-parent households in that city or in each area of the city, for administrative reasons. On the other hand, for scientific purposes we will typically wish to consider the actual population sampled *as if* it was a realisation of a conceptually infinite population extending through time, and also perhaps through space. By taking this latter view, we are able to make generalisations and predictions beyond the units that comprise the real population that has actually been sampled according to a specified sampling scheme.

If we adopt this conceptualisation, often termed a 'superpopulation' or 'model-based' approach, then multilevel modelling becomes a natural way to proceed. Nevertheless, we do need to consider whether the sampling selection process is relevant to the modelling. Generally speaking, if the sampling process (clustering and stratification) is properly incorporated within the model, as described above, we need pay no further attention to the sample selection process, for instance the sampling design weights associated with each unit. Even where we do incorporate terms for stratification and clustering, there may still be an issue as to whether this

has been done adequately, for example, whether all possible interactions and random coefficients are included: we may still wish therefore to incorporate elements of the design into the analysis. Furthermore, we may sometimes wish to make 'marginal' estimates rather than condition, say, on the random effects that reflect the population structure. In this case, we might revert to a weighted analysis that does not include the random effects of a multilevel model.

We shall assume for the present that any sample clustering, where there is variation between clusters, is of interest and thus modelled, although all aspects of the sample design may not be incorporated fully into the model and we may wish to incorporate weights (Chapter 3) for those aspects that are not. We look first at the case where the design is independent of the response variable values. We assume that the sampling scheme generates a set of sample selection probabilities for each unit at each level of the modelled hierarchy. For a 2-level model this will be the set $\{\pi_j, \pi_{i|j}\}$ and this then provides corresponding sampling weights $\{\pi_j^{-1}, \pi_{i|j}^{-1}\}$. We now carry out an analysis using these weights, as described in Section 3.4. Thus, the model estimates take account of the sample design and such a weighted analysis ensures consistent estimates, even where the model does not fully incorporate the sample design, whether by mistake or deliberately.

The second case is where the sample design depends on the values of the response, for example, by stratifying on the basis of the mean value of the response variable for a geographical area, obtained for example from Census data. This is a case of 'informative sampling'. If the stratification is modelled, for example using dummy variables, we will have a model where one or more explanatory variables is correlated with one or more random effects. Another example is where the response is associated with the size of a sampling unit and hence with the selection probability. In their detailed investigation of different weighting procedures, Pfeffermann *et al.* (1998) show that simple procedures involving a scaling of the weights, such as in Section 3.4.1, may not work well and propose a 'psuedolikelihood' estimator instead. Alternatively, the MCMC method described in Section 3.4.2 can be used.

10.2.2 Finite population inference

In the case of finite populations, the traditional approach of 'design-based inference' aims to make inferences about the finite population quantities, for example means or proportions. These may be computed as simple (weighted) functions of the observations with standard errors constructed to reflect the sample design, and we may wish to provide these for the total population or for subpopulations or 'domains'; these domains may also sometimes coincide with clusters or strata. So-called 'model-assisted' inference extends this by using measured 'auxiliary' variables (X) in a regression model. Suppose that these auxiliary variables are available for every population unit and we have a 2-level model where the higher level units are clusters. For the sample we first define the 'multilevel synthetic estimate' using, for example, a simple variance components model such as

$$y_{ij}^{(s)} = (X^{(s)}\beta)_{ij} + u_{0j} + e_{ij} \tag{10.1}$$

where the superscript *(s)* indicates that the observation is a member of the sample. Using the estimated regression coefficients β, we can now form a prediction for every member of the population

$$\hat{y}_{ij} = (X\hat{\beta})_{ij} + \hat{u}_{0j} \tag{10.2}$$

A synthetic estimate is then simply a suitable combination of these, for example the total for domain j is given by $\hat{y}_j = \sum_i \hat{y}_{ij}$. This estimator, however, is 'design-biased', that is it provides biased estimates with respect to repeated sampling from the same finite population. To attempt to correct for this bias a generalised regression (GREG) estimator has been suggested (Sarndal *et al*., 1992). In the present case we can apply the GREG estimator to a multilevel model, hence known as the MGREG estimator (Lehtonen and Veijanen, 1999), to give

$$\hat{y}_j = \sum_{i \in D_j} \hat{y}_{ij} + \sum_{i \in D_j} w_{ij} \left(y_{ij}^{(s)} - \hat{y}_{ij}^{(s)} \right) \tag{10.3}$$

where D_j refers to domain j and the w_{ij} are the sample weights, being the inverse of the selection probabilities within the domain, possibly adjusted for nonresponse. The second term on the right-hand side of (10.3) involves the set of level 1 residuals and adjusts the bias in (10.2), at least approximately.

Clearly, (10.1) and (10.3) can be extended to incorporate random coefficients and more complex multilevel structures. Such models can be expected to be more efficient than non-model-assisted estimators. This approach can incorporate nonlinear models, for example, a logistic model for binary responses; in this case the synthetic estimator consists of predicted probabilities. A further extension is to multivariate responses with appropriate modifications.

10.3 Small area estimation

A feature of many surveys is that they incorporate a large overall sample, but for many domains or areas of interest the numbers are small, so that using just the sample members within such a domain will result in estimates with large standard errors. Thus, a large household survey may select households in a large number of administrative districts, where the sample size within each district is then too small on its own to provide accurate estimates for each district. We shall consider such 'small area estimation' problems for a superpopulation model; for finite populations we can modify the procedures as described above.

The simplest approach to this problem is to fit a multilevel model such as (10.1) and simply estimate the predicted 'synthetic' domain values derived from (10.2). In practice we will often have several surveys sharing some common domains and we look at how these can be used in combination to provide domain estimates that are as efficient as possible.

Suppose we have several data sets $T_1, \ldots T_p$ sharing a common response and also sharing some of the same domains, the higher level area units, for example administrative regions. We shall consider the case of a single response to begin with and consider the extension to multiple responses later. Interest will typically lie in predicting domain means or totals. A special case is where one of the datasets has a special status, for example, a population census or a set of administrative records. We may be able to treat the population estimates from such datasets as covariates in the model for the responses from the other surveys.

We shall use a variance components formulation, but the extension to random coefficients is straightforward, as is the extension to more complex models with further levels and cross classifications.

For dataset T_h we have a 2-level linear variance components model

$$y_{ij}^{(h)} = (X^{(h)}\beta^{(h)})_{ij} + u_j^{(h)} + e_{ij}^{(h)} \tag{10.4}$$

We can allow the fixed and random part explanatory variables to be different in each dataset. Denote the set of domains for the overall model as S_1.

At the domain or area level (2) we have the set of (intercept) random effects

$$\begin{aligned} &u_1^{(1)}, u_2^{(1)} \ldots .. u_1^{(2)}, \ldots \ldots, : \\ &\operatorname{cov}\left(u_j^{(h)}, u_{j'}^{(h')}\right) = 0,\ j \neq j',\ \operatorname{cov}\left(u_j^{(h)}, u_j^{(h')}\right) = \sigma_{u(hh')} \end{aligned} \tag{10.5}$$

By allowing the effects for higher level units to be correlated across surveys, we are able to use all the available information efficiently to provide estimates for each domain. A number of special cases and extensions are of interest.

10.3.1 Information at domain level only

Suppose we have a number of domains (higher level units) where only domain level information is available. Such information might consist of administrative records. A variance components model for this set of domains can be written as

$$y_{.j} = (X\beta)_j + u_j + e_{.j} \tag{10.6}$$

Denote the set of domains for this model by S_2. We have seen how such models can be fitted in Section 3.7. If we now form the union of this set of domains and the previous set, S_1, say $S_3 = \bigcup(S_1, S_2)$, this enhanced set can be analysed as a single model incorporating responses at different levels (Section 3.8.1). A particular case of interest where the inclusion of such data can improve the estimates is for aggregate census data; we note again that the predictors in the component models can be different.

10.3.2 Longitudinal data

Repeated measurements over time can provide information which both increases efficiency and also allows the possibility of estimating trends so that estimates can be updated. Consider a single repeated survey where the same domains are measured across time with different level 1 units, say households, sampled at each occasion. A simple model would consist of a model, for a set of domains S_4, given by

$$\begin{aligned} y_{ijt} &= (X\beta)_{ijt} + \alpha_0 + \alpha_1 t + u_{0j} + u_{1j} t + e_{ijt} \\ e_{ijt} &\overset{iid}{\sim} N(0, \sigma_e^2), \ \begin{pmatrix} u_{0j} \\ u_{1j} \end{pmatrix} \sim N(0, \Omega_u) \end{aligned} \tag{10.7}$$

which incorporates a time trend term in the fixed part and also allows the trend to vary across domains. In some cases, for example, where population census data are available, it may be efficient to condition on the values of the response variable at a prior occasion. In (10.7), both the intercept and slope terms may covary with the domain level random effects for S_3, giving the combined model set $S_5 = \bigcup(S_3, S_4)$, which can be fitted as a single model.

A particular problem with longitudinal data is mobility where individual people, or households, change domains during the course of a study. In principle, such cases can be handled using multiple membership models (Chapter 13). Where time is modelled in such studies, taking account of such mobility also implicitly allows migration patterns to influence the estimates. Where migration is high and 'informative', in the sense that those who move are a nonrandom subsample, taking account of such migration will be important in order to provide unbiased estimates.

10.3.3 Multivariate responses

Models with more than one response variable can be fitted as straightforward extensions to the above models. A particular advantage of such multivariate models is that we can obtain efficient estimates when one of the responses is missing, either completely at random or by design as in the rotation designs considered in Chapter 6. Thus, if information on one response is not included in some datasets by design we can still provide efficient estimates via the correlations with other responses. This will often be a more convenient and flexible method for exploiting these correlations than including the latter responses as covariates. Longford (1999) also addresses this issue and demonstrates the improved efficiency which can result.

11

Multilevel event history and survival models

11.1 Introduction

This class of models has as the response variable the length of time between 'events'. An example is the beginning and end of a period of employment, with the corresponding time being the duration of employment. The term 'survival model' is often used when the end event is terminal, such as a death, in which case repeated episodes are not possible. There is a considerable theoretical and applied literature, especially in the field of biostatistics; a useful summary is given by Clayton (1988) and Singer and Willett (2002) provide illustrative applications. We consider first two basic approaches to the modelling of such duration data. These are based upon 'proportional hazard' models and upon direct modelling of the log duration, often known as 'accelerated life models'. In both cases, we can include explanatory variables. We then discuss discrete (grouped) time models, which are particularly suitable for fitting multilevel data structures.

A multilevel structure in such models can arise in two general ways. The first is where we have repeated durations, or events, within individuals, analogous to our repeated measures models of Chapter 5. Thus, individuals may have repeated employment episodes where they move from one job to another. In this case, we have a 2-level model with individuals at level 2, often referred to as a renewal process. We may also consider more than one *state,* for example, employment and unemployment, with duration models for each state. The second kind of model is where we have a single duration for each individual, but the individuals are grouped into level 2 units. In the case of employment duration, the level 2 units could be firms or employers. If we had repeated events on individuals within firms, then this would give rise

Multilevel Statistical Models: 4th Edition Harvey Goldstein

to a 3-level structure. We can further extend these models where there are cross classifications (Chapter 12) and multiple membership structures (Chapter 13).

11.2 Censoring

A characteristic of duration data is that for some observations we may not know the exact duration but only that it occurred within a certain interval, which is known as interval censored data. Where the start of an interval occurred prior to the observation period, it is known as left censored data, and where it occurred after a known period of observation, for example by an individual leaving a study, it is known as right censored data. For example, if we know at the time of a study only that someone entered her present employment before a certain date, then the information available is left censored and the duration is longer than a known value. We may know that someone entered and then left employment between two measurement occasions, in which case we know only that the duration lies in a known interval. We shall describe procedures for dealing with all kinds of censored data.

In parametric models, where there are relatively large proportions of censored data, the assumed form of the distribution of duration lengths is important, whereas in the semiparametric models the distributional form is ignored. It is assumed that the censoring mechanism is noninformative, namely that the probability of censoring is independent of the actual duration length. We may have data which are censored, but where we have no duration information at all: if we are studying the duration of first marriage for an age cohort of individuals and we end the study when individuals reach the age of 30, all those marrying for the first time after this age will be excluded. To avoid bias, we must therefore ensure that age of marriage is an explanatory variable in the model and report results conditional on age of marriage.

11.3 Hazard and survival functions

The underlying notions are those of *survivor* and *hazard* functions. Consider the (single level) case where we have measures of length of employment on workers in a firm. We define the proportion of the workforce employed for periods greater than t as the *survivor function* and denote it by

$$S(t) = 1 - F(t) = 1 - \int_0^t f(u)du$$

where $f(u)$ is the density function of length of employment. The *hazard* function is defined as

$$h(t) = f(t)/S(t)$$

and represents the instantaneous risk, in effect the (conditional) probability of someone who is employed at time t, ending employment in the next (small) unit interval of time.

It is useful to carry out preliminary analyses of event history data by plotting the observed hazard and survivor functions. This can be done by dividing the time scale into short intervals and computing the empirical hazard for each interval by dividing the number who experience the event during the interval by the number of individuals present at the start of interval i, the 'risk set'. Such hazard estimates are often known as Kaplan-Meier estimates. If there is censoring and we assume that the censoring is equally likely to have occurred at any time during the interval, then we subtract half the number of censored observations in that interval from the denominator. The survival function is simply the proportion who have not experienced the event by the start of interval i, plotted against time. Singer and Willett (2002) provide detailed examples.

The simplest model is one which specifies an exponential distribution for the duration, $f(t) = \lambda e^{-\lambda t}(\mathrm{t} \geq 0)$ which gives $h(t) = \lambda$, so that the hazard rate is constant and $S(t) = e^{-\lambda t}$. In general, however, the hazard rate will change over time and a number of alternative forms have been studied (see, for example, Cox and Oakes, 1984). A common one is based on the assumption of a Weibull distribution, namely

$$f(t) = (\alpha/t)e^{\alpha \ln(t)+\delta}e^{-e^{\alpha \ln(t)+\delta}}$$

or the associated extreme value distribution formed by replacing t by $u = \log_e t$.

Another approach to incorporating time varying hazards is to divide the time scale into a number of intervals within which the hazard rate is assumed constant and dummy variables are used for the time intervals. This may be particularly useful where there are 'natural' units of time, for example, based on menstrual cycles in the analysis of fertility, and this can be extended by classifying units by other factors where time varies over categories (see Blossfeld *et al.*, 2007).

The most widely used models are those known as *proportional* hazards models, and the most common definition is $h(t;\eta) = \lambda(t)e^{\eta}$. The term η denotes a linear function of explanatory variables that we model explicitly in Section 11.5. It is assumed that $\lambda(t)$, the baseline hazard function, depends only on time and that all other variation between units is incorporated into the linear predictor η. The variables in η may depend upon time, and in the multilevel case some of the coefficients will also be random variables.

11.4 Parametric proportional hazard models

Where we have known duration times and right censored data, define the cumulative baseline hazard function $\Lambda(t) = \int_0^t \lambda(u)du$ and a variable w with mean $\mu = \Lambda(t)e^{\eta}$, taking the value one for uncensored and zero for censored data. It can be shown (McCullagh and Nelder, 1989) that the maximum likelihood estimates required are those obtained from a maximum likelihood analysis for this model where w is treated

as a Poisson variable. This computational device leads to the loglinear Poisson model for the q-th observed event

$$\ln(\mu_q) = \ln(\Lambda(t_q)) + \eta_q \tag{11.1}$$

where the term $\Lambda(t_q)$ is treated as an offset, that is, a known function of the linear predictor.

The simplest case is the exponential distribution, for which we have $\Lambda(t) = \lambda t$. Equation (11.1) therefore has an offset $\ln(t_q)$ and the term $\ln(\lambda)$ is incorporated into η. We can model the response as a Poisson count using the procedures of Chapter 4, with coefficients in the linear predictor chosen to be random at levels 2 or above. This approach can be used with other distributions. For the Weibull distribution, of which the exponential is a special case, the proportional hazards model is equivalent to the log duration model with an extreme value distribution and we shall discuss its estimation in Section 11.9.

11.5 The semiparametric Cox model

The most commonly used proportional hazard models are known as semiparametric proportional hazard models. We now look at the multilevel version of the most common of these in more detail.

Consider the 2-level proportional hazard model for the jk-th level 1 unit

$$h_{jk}(t; X_{jk}) = \lambda_k(t) \exp(X_{jk}\beta_k) \tag{11.2}$$

where X_{jk} is the row vector of explanatory variables for the level 1 unit and some or all of the β_k are random at level 2. We adopt the subscripts j,k for levels one and two for reasons which will be apparent below.

We suppose that the times at which a level 1 unit comes to the end of its duration period or 'fails' are ordered and at each of these we consider the total 'risk set'. At failure time t_{jk}, the risk set consists of all the remaining level 1 units. Then the ratio of the hazard for the unit which experiences a failure and the sum of the hazards of the remaining risk set units is

$$\frac{\exp(X_{j'k'}\beta_{k'})}{\sum_{j,k} \exp(X_{jk}\beta_k)},$$

which is simply the probability that the failed unit is the one denoted by j', k' (Cox, 1972). It is assumed that, conditional on the X_{jk}, these probabilities are independent.

Several procedures are available for estimating the parameters of this model (see, for example, Clayton, 1991, 1992). For our purposes it is convenient to adopt the following, which involves fitting a Poisson or equivalent multinomial model of the kind discussed in Chapter 4.

At each failure time $l$we define a response variate for each member of the risk set

$$y_{ijk(l)} = \begin{cases} 1 \text{ if } i \text{ is the observed failure} \\ 0 \text{ if not} \end{cases}$$

where i indexes the members of the risk set, and j,k level 1 and level 2 units. If we think of the basic 2-level model as one of employees within firms, then we now have a 3-level model where each level 2 unit is a particular employee and containing n_{jk} level 1 units where $n_{jk}$is the number of risk sets to which the employee belongs. Level 3 is the firm. The explanatory variables can be defined at any level. They may also vary over time, allowing so-called time varying covariates. A simple variance components model for the expected Poisson count can be written as

$$\mu_{jk(l)} = \exp(\alpha_l + X_{jk}\beta + u_k) \tag{11.3}$$

where there is a 'blocking factor' α_l for each failure time. In fact we do not need generally to fit all these nuisance parameters: instead we can obtain efficient estimates of the model parameters by modelling α_l as a smooth function of the time points, using, say, a low order polynomial or a spline function (Efron, 1988). Alternatively, as suggested in Section 11.3, we can form a piecewise constant hazard model by grouping the blocking factors into relatively homogeneous groups over longer time intervals, possibly based upon a preliminary inspection of the survivor function.

An estimator of the baseline surviving fraction for an individual in the k-th firm at time h, where $X_{jk} = 0$, is

$$\hat{S}_{hk} = \exp\left(-\sum_{l \le h} e^{\hat{\alpha}_l + \hat{u}_k}\right)$$

and the estimate for an individual with specific covariate values X_{jk} is

$$\hat{S}_{hk}^{\{\exp(X_{jk}\beta)\}} \tag{11.4}$$

Where we fit polynomials to the blocking factors, the $\hat{\alpha}_l$ are estimated from the polynomial coefficients, and the surviving fraction can be plotted against the time associated with each interval.

11.6 Tied observations

We have assumed so far that each failure time is associated with a single failure. In practice, many failures will often occur at the same time, within the accuracy of measurement. These are known as 'ties'. Sometimes, data may also be deliberately grouped in time. In this case all the failures at time l have a response $y_{ijk(l)} = 1$. This procedure for handling ties is equivalent to that described by Peto (1972) (see also McCullagh and Nelder, 1989).

11.7 Repeated events proportional hazard models

As with ordinary repeated measures models described in Chapter 5, we can consider the case of multiple episodes, or durations within individuals, with between- and within-individual variation, and possibly further levels where individuals may be nested within firms, etc. The models of previous sections can be applied to such data, but there are further considerations which arise. Where each individual has the same fixed number n of episodes, we can treat these, as in Chapter 6, as constituting n variates so that we have an n-variate model with an $(n \times n)$ covariance matrix between individuals. The variates may be either really distinct measurements or simply the different episodes in a fixed ordering. This is the model considered by Wei *et al*. (1989). We can also model a multivariate structure where, within-individuals, there are repeated episodes for a number of different types of event. For each type we may have coefficients random at the individual level and these coefficients will generally also covary at that level.

Often with repeated events models the first episode is different in nature from subsequent ones. An example might be the first episode of a disease which may tend to be longer or shorter, and with different determinants, than subsequent episodes. If the first episode is treated as if it were a separate variate, subsequent episodes can be regarded as having the same distribution, as in the previous section. Another example is the length of the interval to a first birth from the start of a marriage or partnership. We give an example of such a model later.

Another possible complication in repeated event data, as in Chapter 5, is that we may not be able to assume independence between durations *within*-individuals. This will then lead to serial correlation models which can be estimated using the procedures discussed in Chapter 17, for the parametric log duration models discussed below. We shall return to a further exploration of repeated events models when we look at discrete time models.

11.8 Example using birth interval data

The data are a series of repeated birth intervals for 379 Hutterite women living in North America (Larsen and Vaupel, 1993; Egger, 1992). The response is the length of time in months from birth to conception of next child, ranging from 1 to 160, with the first birth interval ignored and no censored information. This gives 2235 births in all.

There is information available on the mother's birth year, her age in years at the start of the birth interval, whether the previous child died within one year of birth, and the duration of marriage at the start of the birth interval. Table 11.1 gives the results from fitting a proportional hazards model using the formulation in (11.5) with random coefficients for the intercept and whether the previous child died. A fourth order polynomial was adequate to smooth the blocking factors.

The only coefficient estimated with a nonzero variance at level 2 was whether or not the previous birth died, but a large sample chi-bar test (Section 2.8), on 2 degrees

Table 11.1 Proportional hazards model for Hutterite birth intervals. In the random part, subscript 0 refers to intercept, 1 to previous death. PQL2 estimates.

Parameter	Estimate (s.e.)	Estimate (s.e.)
Fixed	Model A	Model B
Intercept	−3.65	−3.64
Mother's birth year – 1900	0.026 (0.003)	0.026 (0.003)
Mother's age (year – 20)	−0.008 (0.014)	−0.004 (0.014)
Previous child died	0.520 (0.118)	0.645 (0.144)
Marriage duration (months)	−0.003 (0.001)	−0.004 (0.001)
Random		
σ^2_{u0}	0.188 (0.028)	0.188 (0.028)
σ_{u01}		0.005 (0.088)
σ^2_{u1}		0.381 (0.236)

of freedom, for the two random parameters for this coefficient gives a P-value of 0.005. An increase in the linear predictor is associated with a shorter interval. Thus the birth interval is shorter for the later born mothers and also if the previous child died. The interval is somewhat longer the longer the marriage duration, with little additional effect of maternal age. This apparent lack of a substantial age effect seems to be a consequence of the high correlation (0.93) between duration of marriage and age. Higher order terms for duration and age were fitted, but the estimated coefficients were small and not significant at the 10 % level. The between-individual standard deviation in Model B is about 0.4, which is comparable in size to the effect of a previous death. The between-individual standard deviation for a model with just an intercept is 0.45 so that the covariates explain only a small proportion of the between-individual variation. Figure 11.1 shows two average estimated surviving fraction curves for a woman aged 20, born in 1900 with marriage duration 12 months. These curves represent the median, rather than mean, surviving fractions (at a random effect value of zero) since we are here dealing with a nonlinear model. The higher one is for those where there was a previous live birth and the lower where there was a previous death.

11.9 Log duration models

For the accelerated life model the distribution function for duration is commonly assumed to take the form

$$f(t; X, \beta) = f_0(te^{X\beta})e^{X\beta}$$

where f_0 is a baseline function (Cox and Oakes, 1984).

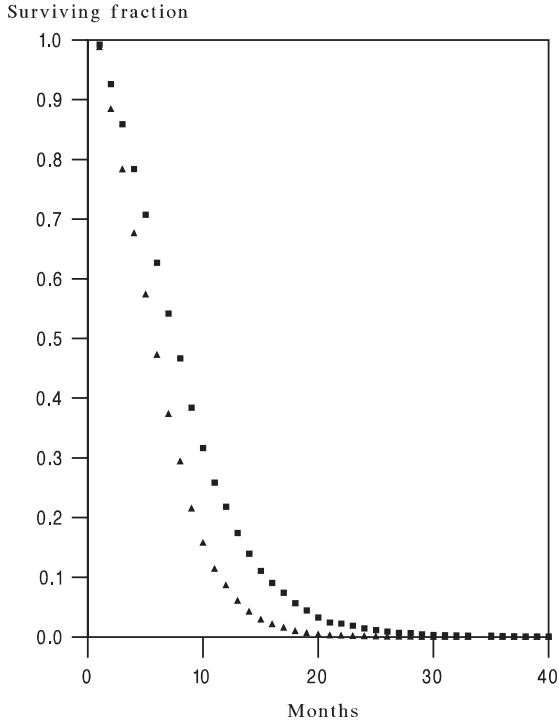


Figure 11.1 Probability of birth after month t last born alive upper, last born died within one year lower.

For a 2-level model we can write the log duration

$$l_{jk} = \log_e(t_{jk}) = X_{jk}\beta_k + e_{jk} \tag{11.5}$$

which is in the standard form for a 2-level model. We shall assume normality for the random coefficients at level 2 (and higher levels) but at level 1 we shall study other distributional forms for the e_{ij}. The level 1 distributional form is important where there are censored observations. We first consider the common choice of an extreme value distribution for the log duration distribution, L, conditional on $X_{jk}\beta_k$, which implies an equivalence with the proportional hazards model. Omitting level subscripts we write

$$f(l;\alpha,\delta) = \alpha e^{-\alpha l+\delta}\exp(-e^{-al+\delta}) \quad -\infty < l < \infty \tag{11.6}$$

$$E(L) = \alpha^{-1}(\delta+\gamma), \quad \text{var}(L) = \frac{\pi^2}{6\alpha^2}, \quad \gamma = 0.5772$$

For (11.5) this gives

$$\begin{aligned}\pi_{jk} &= \Pr(L > l_{jk}) = 1 - \exp(-e^{-\alpha l_{jk}+\delta_{jk}}) \\ \pi'_{jk} &= \alpha.\exp\{-e^{-\alpha l_{jk}+\delta_{jk}}\}e^{-\alpha l_{jk}+\delta_{jk}}\end{aligned} \tag{11.7}$$

where the differential is for use in the estimation of censored data and is with respect to β in the expression below.

The mean of L is incorporated into the fixed predictor. If we have no censored data, we estimate the parameters for the model given by (11.5) by treating it as a standard multilevel model. We note that the estimation is strictly quasilikelihood, since we are using only the mean and variance properties of the level 1 distribution. If we assume a simple level 1 variance, then we can iteratively estimate α from the above relationship and we also obtain for the 2-level model (11.5)

$$\delta_{jk} = \gamma + \alpha(X_{jk}\beta_k)$$

where there is complex variation at level 1 then α will vary with the level 1 units. To estimate the survival function for a given level 2 unit, we first condition on the covariates and random coefficients, that is $X_{jk}\beta_k$, and then use (11.7).

We can choose other distributional forms for the log duration distribution. These include the log Gamma distribution, the normal and the logistic. Thus, for example, for the normal distribution we have

$$\begin{aligned}\pi_{jk} &= 1 - \Phi(z_{jk}) \\ \pi'_{jk} &= \phi(z_{jk})/\sigma_e \\ z_{jk} &= [l_{jk} - (X\beta)_{jk}]/\sigma_e\end{aligned}$$

where Φ, ϕ are the cumulative and density functions of the standard normal distribution. Quasilikelihood estimates can be obtained for any suitable distribution with two parameters. The possibility of fitting complex variation at level 1 can be expected to provide sufficient flexibility using these distributions for most purposes.

11.9.1 Censored data

Where data are censored in log duration models we require the corresponding censoring probabilities. Thus, for right censored data we would use (11.7) with corresponding formulae for interval or left censored data. For each censored observation we therefore have an associated probability, say π_{ij} with the response variable value of one.

This leads to a bivariate model in which for each level 1 unit the first response is the continuous log duration time if not censored; the second takes the value 1 if censored and zero if not, with corresponding explanatory variables in each case. There are basically two sets of explanatory variables for the level 1 variation, one for the continuous log duration response and one for the binary response. In the former case, we can extend this for complex level 1 variation, as in the example analysis

below. For the latter, we can use the standard logit model, as described in Chapter 4, possibly allowing for extrabinomial variation. The random parameters at level 1 for the two components are uncorrelated. When carrying out the computations, we may obtain starting values for the parameters using just the uncensored observations.

Since the same linear function of the explanatory variables enters into both the linear and nonlinear parts of this model, we require only a single set of fixed part explanatory variables, although these will require the appropriate transformation for the logit response, as described in Chapter 4. We also note that any kinds of censored data can be modelled, so long as the corresponding probabilities can be modelled.

We can readily extend this model to the multivariate case where several kinds of durations are measured. This will require one extra lowest level to be inserted to describe the multivariate structure, with level 2 becoming the between-observation level and level 3 the original level 2. For the binary part of the model we will allow correlations at level 2 where these can be interpreted as point-biserial correlations. If we fit a probit model, as discussed in Appendix 4.3, we can interpret such correlations on a continuous scale and this will often be more satisfactory.

For repeated events models where there are different types of duration, we can choose to fit a multivariate model. Alternatively, we may be able to specify a simpler model where the types differ only in terms of a fixed part contribution, or perhaps where there are different variances for each type with a common covariance. As pointed out earlier, we may sometimes wish to treat the first duration length separately and this is readily done by specifying it as a separate response.

11.9.2 Infinit durations

It is sometimes found that for a proportion of individuals, their duration lengths are extremely long or effectively infinite. Thus, some employees remain in the same job for life and some patients may acquire a disease and retain it for the rest of their lives. In studies of social mobility, some individuals will remain in a particular social group for a finite length of time while others may never leave it: such individuals we can refer to as 'long term survivors'. We will treat such durations as if they were infinite. Since any given study will last only for a finite time, it is impossible precisely to distinguish long term survivors from those that are right censored. Nevertheless, if we make suitable distributional assumptions we can obtain an estimate of the proportion of such survivors, θ.

For a constant θ, given an incompletely observed duration time, the observation is either right censored with finite duration or has infinite duration so that we replace the probability π_{ij} by $\lambda_{ij} = (1-\theta)\pi_{ij} + \theta$. In general, θ will depend on explanatory variables and an obvious choice for a model is

$$\text{logit}(\theta_{ij}) = X_{ij}^{(\theta)}\beta^{(\theta)}. \tag{11.8}$$

Some of the coefficients in (11.8) may also vary across level 2 units.

Where the observation is not censored we know that it has a finite duration, so that for an unobserved duration time we have a response variable taking the value zero with predictor given by $\{1 + \exp -(1 - \theta_{ij})\}^{-1}$. The full model can therefore be specified as a bivariate model where for observed durations we have two responses, one for the uncensored component l_{ij} and the one for the parameters $\beta^{(\theta)}$. For the incompletely observed observations there is a single response which takes the value one with predictor function

$$\{1 + \exp -[(1 - \theta_{ij})\pi_{ij} + \theta_{ij}]\}^{-1}.$$

We can extend the procedures of Chapter 4 to the joint estimation of β, $\beta^{(\theta)}$, noting that for the censored observations when estimating β, we have

$$\lambda'_{ij}(\beta) = (1 - \theta_{ij})\pi'_{ij}$$

and for estimating $\beta^{(\theta)}$ we have

$$\lambda'_{ij}(\beta^{(\theta)}) = (1 - \pi_{ij})\theta'$$

11.10 Examples with birth interval data and children's activity episodes

We first look again at the Hutterite birth interval data. Since all the durations are uncensored, we apply a standard model to the log(birth interval) values. Results are given in Table 11.2.

We see that we can now fit the year of birth and age as random coefficients at level 2. We have significant heterogeneity at level 1 where there is a greater within woman variation in duration of intervals following a child death, with a chi-bar on 1 d.f. of 4.7, P = 0.02. As before, mother's birth year and previous death are associated with a decrease and duration of marriage with an increase in birth interval. Note that the estimated surviving fraction will in general depend on the level 1 distributional assumption.

In the present case, as shown in Figure 11.2, the level 1 standardised residuals show little departure from normality; Figure 11.3 shows the estimated surviving fraction, based on normality, for women born in 1900, with marriage duration 12 months, aged 20 and with a previous survived birth.

Figure 11.3 is similar to Figure 11.1, based on the proportional hazards model. In fact, the two lines actually cross at about 30 months, as a result of the different level 1 variances for those with a previous survived birth as opposed to a death.

We now look at some data which exhibit more extensive variance heterogeneity at level 1. They measure the number of days spent by pre-school children either at home or in one of six different kinds of pre-school facility. For each of 249 children

Table 11.2 Log duration of birth interval for Hutterite women. For the random parameters subscript 0 refers to the intercept, 1 refers to birth year, 2 to age and 3 to previous death.

Parameter	Estimate (s.e.)	Estimate (s.e.)	Estimate (s.e.)
Fixed	A	B	C
Intercept	1.97	1.96	1.97
Mother's birth year – 1900	−0.021 (0.002)	−0.021 (0.002)	−0.021 (0.002)
Mother's age – 20	−0.005 (0.010)	−0.005 (0.010)	−0.005 (0.010)
Previous death	−0.435 (0.079)	0.436 (0.079)	−0.438 (0.089)
Marriage duration (months)	0.003 (0.001)	0.003 (0.001)	0.003 (0.001)
Random			
Level 2			
σ^2_{u0}	0.127 (0.017)	0.114 (0.052)	0.121 (0.054)
σ_{u01}		−0.001 (0.002)	−0.001 (0.002)
σ^2_{u1}		0.0001 (0.0001)	0.0001 (0.0001)
σ_{u02}		−0.004 (0.003)	−0.005 (0.003)
σ_{u12}		0.0001 (0.0001)	0.0001 (0.0001)
σ^2_{u2}		0.0005 (0.0003)	0.0006 (0.0003)
Level 1			
σ^2_{e0}	0.549 (0.018)	0.533(0.018)	0.522 (0.018)
σ^2_{e3}			0.200 (0.108)
−2 loglikelihood	5305.9	5295.5	5290.8

there were up to 12 periods of activity. Further details can be found in Plewis (1985, Chapter 7).

The response is the logarithm of the number of days and covariates are the type of episode, with home chosen as the base category and the education of the mother measured on a 7-point scale ranging from no education beyond minimum school leaving age (0) to university degree (6). Nineteen of the episodes were right censored and 25 were left censored, being less than one day.

The multilevel structure is that of episodes within children. The model is also multivariate with the type of activity as six response variables, covarying at the level of the child. Table 11.3 shows the results of an analysis where there is a single between-child variance and where it is allowed to differ for each type of episode. The between-episode-within-child variance is also allowed to vary for different episodes. The level 1 residuals for the continuous response part of the model show some evidence of non normality and we therefore show the results for the extreme value distribution. Because of the relatively small amount of censoring there is little difference for the parameter estimates between analyses making different distributional assumptions.

We see that there is quite substantial variation at both levels. At level 2 there was between-children variation only for facility types 1,2 and 4. A proportional hazards

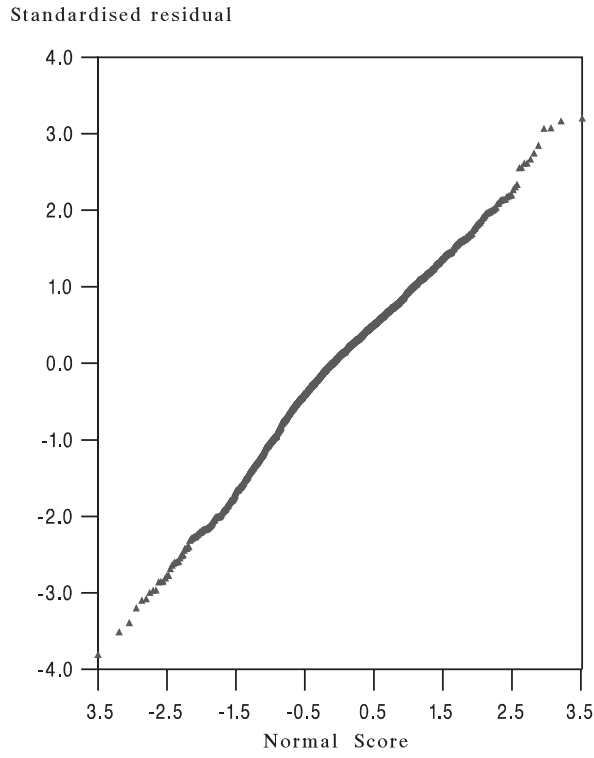


Figure 11.2 Level 1 residuals by Normal scores for Analysis B in Table 11.2

model fitted to these data did not show any between-child variation. In general, the semiparametric proportional hazards model will not detect some of the relationships apparent from fitting parametric models although it has the advantage that it does not make strong distributional assumptions. Figure 11.4 shows the predicted probabilities of home and facility type 1 episodes lasting beyond various times expressed in log (days). The crossing of the lines is now much clearer as a consequence of the different level 1 variances.

11.11 The grouped discrete time hazards model

Where time is grouped into preassigned categories, we write the survivor function at start of time interval t, the probability that failure occurs after the start of this interval, as s_t. This gives

$$f_t = s_{t-1} - s_t, \quad h_t = f_t/s_{t-1}, \quad s_0 = 1$$

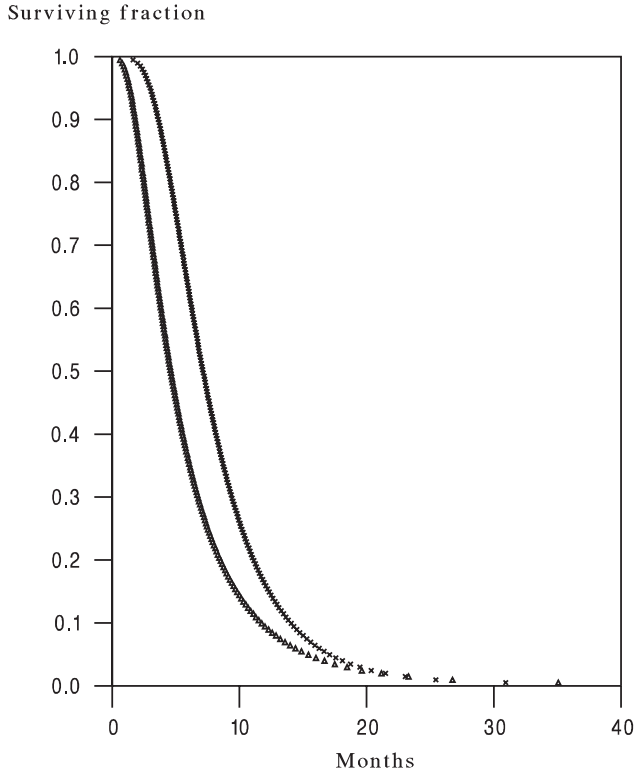


Figure 11.3 Estimated survival functions for women with previous live births (upper) and a previous death; born in 1900, age 20, 12 months marriage.

and

$$s_t = \prod_{l=1}^{t} (1 - h_l)$$

which can be used to estimate the survivor function from a set of estimated hazards. Thus, the basic record is one record for each time interval within each higher level unit, with the response being a binary indicator of failure for each interval. The estimation follows that for the binary response model (Chapter 4) and we can use the common choice of a logit link function, although other links such as the probit are possible. (For example, Aitkin *et al*., 1989, discuss a log-log link leading to a proportional hazards model.) Censored observations are simply excluded from the relevant set.

As in the semiparametric Cox model, we can fit a polynomial function to the successive time intervals, rather than the full set of blocking factors. The data will be ordered within level 2 units so that a risk set in general will extend over several

Table 11.3 Log duration analysis of children's activity episodes: extreme value distribution.

Parameter			
Fixed	A (s.e.)	B (s.e.)	
Intercept	2.19	2.18	
Facility 1	−0.12 (0.11)	−0.13 (0.11)	
Facility 2	0.20 (0.08)	0.18 (0.08)	
Facility 3	0.00 (0.13)	0.00 (0.13)	
Facility 4	0.87 (0.12)	0.95 (0.11)	
Facility 5	0.28 (0.09)	0.28 (0.09)	
Facility 6	0.15 (0.09)	0.14 (0.08)	
Mother's education	−0.05 (0.02)	−0.05 (0.02)	
Random			
Level 1 variance			
Overall	0.75		
Home		0.76	
Facility 1		1.23	
Facility 2		0.83	
Facility 3		0.79	
Facility 4		0.40	
Facility 5		0.65	
Facility 6		0.57	
Level 2 covariance matrix. Analysis A (Analysis B in brackets)			
	Facility 1	Facility 2	Facility 4
Facility 1	0.34 (0.0)		
Facility 2	0.11 (0.0)	0.20 (0.17)	
Facility 4	−0.28 (0.0)	0.13 (0.09)	0.07 (0.23)

such units. A general procedure is to specify the response for each level 1 unit as binary, that is zero if the unit survives the interval and one if not. Thus, a 2-level model will become specified as a 3-level model with the binomial variation at level 1 and the actual level 1 units at level 2. This will be necessary if there are time varying covariates.

We now consider various extensions to this model to incorporate multivariate sequences, competing risks and multiple starting states. It is convenient to discuss these models for a 2-level repeated events structure where each individual has a sequence of event episodes. Finally, we consider a discrete time model where the set of intervals is treated as an ordered response and show how this can accommodate models with mixed multivariate responses and missing data.

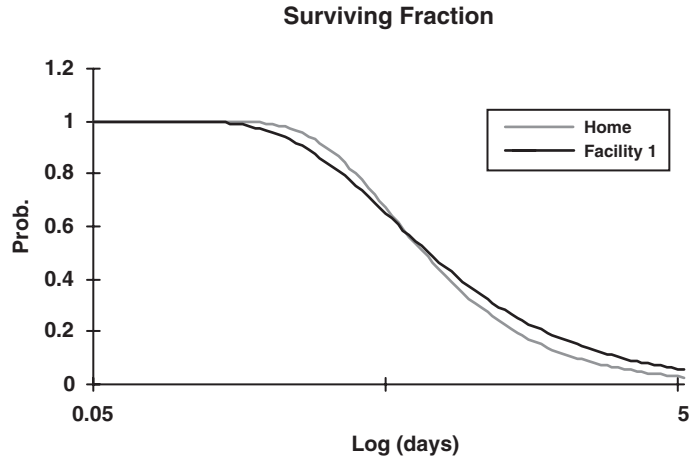


Figure 11.4 Estimated surviving probability of activity episodes.

11.11.1 A 2-level discrete time event history model for repeated events

We shall discuss this model using an example from the National Child Development Study of co-residential partnership durations (Bynner *et al*., 2002). In this model we have repeated events for each of two states – not being in a partnership (whether married or not) and belonging to a partnership. We assume that the total time interval is divided into short time intervals, for example, three months, within which at most one transition is assumed to have taken place per individual. These actual time intervals are recoded into *modelled time intervals* (z) grouped within an episode (k), determined by the event state. Data for individual 1 may look as in Table 11.4.

Thus, starting in state 'no partnership' (episode 1), individual 1 moves in time interval 3 to state 'partnership' (episode 2) and in time interval 7 to state 'no partnership' (episode 3). The response variable, y, takes the value zero if no move takes place and one if a change in partnership status occurs during the interval. Thus the hazard at modelled time t is

$$h_{ijk(t)} = P(y_{ijk(t)} = 1 | y_{ijk(t-1)} = 0)$$

where k indexes individual, j indexes episode and i indexes the state. The states (partnership, non-partnership) are modelled by dummy variables and this leads to a binary response model where, as we have assumed, the (conditional) responses within episodes within individuals are independent. Level 3 is the individual, level 2 represents variation between repeated episodes within individuals and level 1 refers to the time interval within repeated episodes. More generally, we can consider models

Table 11.4 Discrete time interval data for partnership duration.

Individual	Actual time interval	Modelled time interval	Response	Event state at start of interval
1	1	1	0	no partnership
1	2	2	0	no partnership
1	3	3	1	no partnership
1	4	1	0	partnership
1	5	2	0	partnership
1	6	3	0	partnership
1	7	4	1	partnership
1	8	1	0	no partnership
1	9	2	0	no partnership
1	10	3	0	no partnership

with several possible states through which individuals move, with individual level correlated random effects.

Using a logit link function, this model can be written in the form

$$\begin{aligned} \text{logit}(h_{ijk(t)}) &= \beta_0 + \sum_{h=1}^{p} \alpha_h (z_{i(t)})^h + \sum_{j=1}^{m} \beta_l x_{lijk(t)} + u_{ijk} + v_{ik} \\ y_{ijk(t)} &\sim bin(1.h_{ijk(t)}) \end{aligned} \tag{11.9}$$

where $z_{i(t)}$ indexes the modelled interval at discrete time t using a p-order polynomial (typically $p \leq 5$) to describe the baseline hazard (Section 11.5), and $x_{kij(t)}$ are covariates, including the dummy variable for partnership state and its interactions with the other covariates to allow for different prediction functions for the two states. The term v_{ik} is the random effect for individual k for state i and u_{ijk} is a random effect associated with the j-th episode for the k-th individual. In fact the term u_{ijk} represents extra-binomial variation within individuals across repeated episodes. In a 2-level model with just one episode per individual this is often known as a 'frailty' term.

We retain the subscript j in the covariate expression to allow for episode or time varying covariates. Where there are many discrete time intervals this avoids the estimation of a large number of nuisance terms; one for each ordered time interval. At the individual level (2) we may have several random effects, in particular for each state, and all of these will covary.

The population probability of survival to the end of modelled time interval t for state i is

$$\prod_{m=1}^{t} (1 - h_{ijk(m)}) \tag{11.10}$$

where these can be averaged over individual random effects to provide population estimates.

11.11.2 Partnership data example

The National Child Development Study (NCDS) is a longitudinal study which takes as its subjects all those living in Great Britain who were born between 3 and 9 March 1958. The fifth follow-up took place when they were 33. A self-completion questionnaire asked for retrospective information on relationships, children, jobs and housing from the age of 16. Altogether, 11178 persons responded and all but 39 of the cohort members had no more than four partnerships.

We confine ourselves to male cohort members and episodes that start with a partnership because the first non-partnership episode starting at age 16 is untypical (but see Goldstein *et al*., 2004). At level 2 we have two random effects, one for partnership and one for non-partnership durations. The total number of three-month periods is 140 420 with 3737 male cohort members who had at least one partnership.

The MCMC estimation for (11.9) was run for 10 000 iterations with a burn in of 1000. It gives somewhat different estimates from PQL1, as noted in Chapter 4. From Table 11.5 we see that there is some evidence that those with manual occupations have longer partnership durations (the negative coefficient is associated with a lower probability of terminating an episode at any given time), but a small and non-significant difference ($-0.116 + 0.152 = 0.036$) for the non-partnership durations. The later the starting age the longer the duration for partnerships but there is only a small ($-0.063 + 0.049 = -0.014$) and non-significant relationship for non-partnerships. We note also that there is a negative correlation of –0.52 between partnership and non-partnership episodes indicating that long partnerships tend to be associated with short non-partnerships and vice versa. Individuals can tentatively be classified on this basis as either long partnership/ short non-partnership or long non-partnership/ short partnership individuals.

11.11.3 General discrete time event history models

We now look at some extensions to the model. We write (11.9) in the general form for a 2-level model

$$\begin{aligned} &\text{logit}(h_{ij(t)}) = (Z\alpha)_{it} + (X\beta)_{ij(t)} + u_{ij} \\ &y_{ij(t)} \sim Bin(1, h_{ij(t)}), \quad u_{ij} \sim MVN(0, \Omega_u) \end{aligned} \tag{11.11}$$

We note that (11.11) allows time varying covariates. In general, we can model a cohort of individuals moving through time where the response indicates whether an event occurs during interval t. For simplicity, we assume in the following discussion that the observations are for intervals within individuals, as in the example of the previous section.

A *competing risks* model is one where there are several ways for an interval to end, in addition to several ways for states to start, indexed by i. Suppose that there

Table 11.5 Random coefficient model for partnership and outside partnership. MCMC estimates: starting from first partnership.

Parameter	Estimate (s.e.)
Fixed	
intercept	−3.709
z	$-0.244(0.067)^{*}10^{-1}$
z^2	$0.076(0.053)^{*}10^{-2}$
z^3	$-0.140(0.240)^{*}10^{-4}$
z^4	$-0.089(0.114)^{*}10^{-5}$
z^5	$0.093(0.264)^{*}10^{-7}$
start age	−0.063(0.010)
manual	−0.116(0.075)
np(non-partnership)	1.753(0.469)
np*z	$0.499(0.208)^{*}10^{-1}$
np*z^2	$-0.181(0.134)^{*}10^{-2}$
np* z^3	$-0.112(0.114)^{*}10^{-3}$
np* z^4	$0.024(0.332)^{*}10^{-5}$
np* z^5	$0.055(0.158)^{*}10^{-6}$
np*start age	0.049(0.017)
np*manual	0.152(0.118)
Random	
σ^2_{v0}(non partnership)	0.400(0.211)
σ_{v01}	−0.119(0.118)
σ^2_{v1}(partnership)	1.145(0.171

are $R-1$ end events. Denote the multinomial response by $y_{ij(t)}$ where $y_{ij(t)} = r$ if an event of type r has occurred in time interval t, $r = 2, \ldots, R$, and $y_{ij(t)} = 1$ if no event has occurred. The hazard of an event of type r in interval t, denoted by $h^{(r)}_{ij(t)}$, is the probability that an event of type r occurs in interval t, for state i given that no event of any type has occurred before interval t. The log-odds of an event of type r versus no event is modelled using a multinomial logit model as defined in Chapter 4 as follows (see Steele *et al*., 1996, who introduce this model).

$$\log\left[\frac{h^{(r)}_{ij(t)}}{h^{(1)}_{ij(t)}}\right] = (Z\alpha)^{(r)}_{(t)} + (X\beta)^{(r)}_{ij(t)} + u^{(r)}_{ij} \quad r = 2, \ldots, R. \tag{11.12}$$

We may also have different sets of end events for each state in which case (11.12) is modified so that there are R_i end events for state i. In a competing risks model, the effects of duration and covariates may differ for each event type, as indicated by the r superscript. It is also possible that the form of the baseline hazard and the

set of covariates may vary across event types. The random effects are assumed to follow a multivariate normal distribution, with covariance matrix Ω_u, and we will in general have an effect for each combination of state and end event. A simplification will occur if we are prepared to assume that, for each end event, there is a random effect irrespective of initial state so that (11.12) becomes

$$\log\left[\frac{h^{(r)}_{ij(t)}}{h^{(1)}_{ij(t)}}\right] = (Z\alpha)^{(r)}_{(t)} + (X\beta)^{(r)}_{ij(t)} + u^{(r)}_i + u^{(r)}_j \tag{11.13}$$

which has the form of a cross-classified model (Chapter 12).

We can fit all of these models using various link functions such as thelogit. In the case of the probit link, we can interpret the hazard as the cumulative probability for an underlying standard normal distribution. Thus, for a given time period, covariates (and random effects) corresponding to an observed response, we can estimate the value on the underlying normal density, interpreted as a propensity to end an episode at time t (see Chapter 4). The values of these at any time and set of covariate values therefore provides an alternative interpretation for the model.

Where the transition states form an ordered categorisation, we can use corresponding ordered category models for this, for example, by modelling cumulative log-odds (Chapter 4). This could arise in the modelling of illness duration, where patients make transitions between clinical states which are ordered by severity. For such models we might also assume an underlying propensity with a probit link.

These models can be extended to the multivariate case where, for each individual, we wish to study more than one type of episode at a time; for example, duration of partnership episodes and duration of employment episodes. For each episode type we form the same set of discrete elementary time intervals and the response is a p-way table where p is the number of episode types. This table is treated as a multinomial response with corresponding, dummy, explanatory variables for the margins of the table that we wish to fit. If covariates are present, these will be modelled by interacting them with the explanatory variable dummies. For ordered models and for binary response models with a probit link function, we can directly incorporate correlations between the underlying normal distributions at level 1. This then provides covariance matrix estimates for the different episode types at all levels of the data hierarchy and allows us to study the relationship between the different types. Such a model is also referred to as a multiprocess model.

Discrete time models provide a very flexible means of modelling multilevel event history data since they just involve the use of existing procedures for discrete responses. One disadvantage of the formulation that we have been considering is that it requires data in a form that uses large amounts of memory, and in the next section we show how this can be avoided with an alternative formulation. Steele *et al*. (2004) discuss these models in more detail, with examples, and show how they can be fitted with existing software.

11.12 Discrete time latent normal event history models

We now consider treating event duration as an ordered categorical variable and applying the range of procedures developed in Chapter 7. This has certain advantages.

Suppose $f(t > 0)$ is a survival or event time distribution. We can transform to a standard normal distribution by defining the cumulative distribution as

$$F(x) = pr(0 < t \leq x) = \int_{-\infty}^{x} \phi(z)dz.$$

One way to operationalise this model is to categorise the time scale by defining cut points ($t_0(= 0), t_1, \ldots .t_p$) and to consider the cumulative distribution $\Pr(t \leq t_h)$. We allow $t_p = \infty$. This thus defines the ordered probit model of Chapter 7, which can be written, ignoring subscripts, as

$$\begin{aligned} &\gamma^{(h)} = \int_{-\infty}^{\alpha_h - X\beta} \phi(z)dz, \quad \gamma^{(h)} = \sum_{g=1}^{h} \pi_g \\ &\pi_g = pr(t_{g-1} < t \leq t_g), \quad \alpha_1 = 0 \end{aligned} \tag{11.14}$$

where $X\beta$ represents the effect of any covariates, and π_g is the probability that an event occurs in time interval g. The threshold parameters α_h correspond to the cut points and we require that they satisfy the order constraint $\alpha_1 < \alpha_2 \ldots\ldots < \alpha_p$. We can set $\alpha_1 = 0$ if we assume that the intercept is incorporated in $X\beta$. We generalise to the 2-level case by adding random effects replacing $X\beta$ by $X\beta + Zu, u \sim N(0, \Omega_u)$ with further levels or classifications similarly specified. The hazard for time interval g is $\pi_g/(1 - \gamma^{(g-1)})$ and the hazard *rate* at the start of time interval g is $\phi'(\alpha_{(g-1)} - X\beta)/(1 - \gamma^{(g-1)})$.

An important advantage of this formulation is that the level 1 unit is a single duration episode that records the category where a failure occurs, so that we do not need to expand the dataset. The MCMC algorithm (Chapter 7) samples values from the latent normal distribution, so that we can have multivariate models with different kinds of (correlated) survival times, such as length of time in employment and length of partnership as well as further responses of various types at the individual level such as health status. We can also extend to the case where there are responses at several levels (see Chapter 16).

One drawback of this formulation is that it does not lend itself directly to the incorporation of time varying covariates, but we shall describe how this can be done below where we also reformulate it in a nonlinear form. Before we do this, we describe how the model can deal with right, left and interval censored data, how we can model missing data and how we can incorporate the timing of events within an interval. We note that there is no requirement to have equal intervals on the time scale, and there is no assumption that the hazard is constant within a time interval, although some

information will typically be lost if the hazard does change within an interval. A fully detailed description of this model is given in Goldstein (2010a).

11.12.1 Censored data

If we have right censored data, this implies that $t > t_k$ for some threshold k. Now $\Pr(t > t_k) = \int_{\alpha_k - X\beta}^{\infty} \phi(z)dz$ so that in an MCMC step we sample from the corresponding tail of the standard normal distribution. If right censoring is known to occur during the first interval then we sample from the full normal distribution. In the case of left censored data, we observe when failure occurs but we do not know when the individual became exposed to the risk of failure, so that this gives us information that the duration is longer than the time between when observations began and when the event is observed. We therefore sample as for right censoring. Interval censored data are those that occur in two or more contiguous intervals. In that case, we sample from the union of this set of intervals. When sampling the threshold parameters in an MH step, we include interval censored data using the probability of falling in the corresponding set of intervals.

11.12.2 Missing data

Censored data are an example of missing data where we sample from the corresponding interval or tail of the underlying normal, as described above. We condition on the predictors included in the model and make the standard assumption that data are missing (conditionally) at random (MAR), that is, that the probability of being censored is independent of the actual survival time. There are situations, however, where these assumptions may be violated. In that case there may be further, auxiliary, variables so that if we were to condition on them, we would satisfy the MAR assumption; for example, if we have information on a subject's social environment. In instances where covariates in the model have missing values, a standard procedure is that of multiple imputation (MI); the methods for this, as discussed in Chapter 16, can be applied directly to the ordered survival data to 'impute' the latent normal variables for our model, using multiply imputed data sets for inference.

11.12.3 Information about the timing of events in an interval

As already noted, we may lose useful information through a coarse grouping using a particular set of cut points. Within any interval, however, for those who do not survive to the next time interval, we may have more detailed information on when during the interval an individual has the event. Suppose we divide the grouped intervals into narrower sub-intervals. Denote the start and end times of a particular sub-interval by $t_1(h), t_2(h)$ for the h-th interval ($h = 1, \ldots p-1$). Then instead of sampling a normal variable in the interval $(\alpha_{h-1} - X\beta, \alpha_{h-1} - X\beta)$, we sample in the sub-interval $\alpha_{h-1} + w_1(h) - X\beta, \alpha_{h-1} + w_2(h) - X\beta$ where $w_1(h), w_2(h)$ are

defined by

$$\frac{\int_{\alpha_{h-1}-X\beta}^{\alpha_{h-1}+w_1(h)-X\beta}\phi(z)dz}{\int_{\alpha_{h-1}-X\beta}^{\alpha_{h-1}-X\beta}\phi(z)dz}=\frac{t_{1(h)}-t_{h-1}}{t_h-t_{h-1}},\quad \frac{\int_{\alpha_{h-1}-X\beta}^{\alpha_{h-1}+w_2(h)-X\beta}\phi(z)dz}{\int_{\alpha_{h-1}-X\beta}^{\alpha_{h-1}-X\beta}\phi(z)dz}=\frac{t_{2(h)}-t_{h-1}}{t_h-t_{h-1}} \tag{11.15}$$

where for $h=1$ we set $\alpha_0=-\infty$.

This sub-interval information is ignored when using MH sampling for the threshold parameters. We note that the sub-intervals may differ across individuals; for example, for some individuals or groups of individuals we may have more precise timing information than for others. The assumption in (11.15) is that the probability of a failure occurring between the start of the interval and $w_1(h)$, divided by the probability of an event occurring in any part of the interval, is equal to the time from the start of the interval to $w_1(h)$ divided by the length of the interval; likewise for $w_2(h)$. Clearly, other assumptions are possible but we shall not pursue them.

11.12.4 Modelling the threshold parameters and time varying covariates

We now look at models where we allow the threshold parameters to be modelled as functions of explanatory variables. If these are defined at the individual level, we can incorporate them in $X\beta$ so that we shall consider here models where these explanatory variables are defined at the threshold level, and thus become time varying covariates.

One possibility is to model the threshold parameters as a linear function of explanatory variables. A problem with such a formulation is that at each step the composite threshold parameters must satisfy the relevant monotonicity order restrictions and with time varying covariates this requires testing the condition for each threshold for each individual, and may give rise to poorly mixing MCMC chains. Furthermore, even where such monotonicity restrictions are obeyed for the data set being analysed, this does not guarantee that for some values of the covariates in future data sets these constraints will not be violated. We therefore consider an alternative nonlinear formulation that will satisfy strict monotonicity in general. We reformulate model (11.14) as follows:

$$\begin{aligned}\gamma^{(h)}&=\int_{-\infty}^{\alpha_h-X\beta}\phi(z)dz,\quad \gamma^{(h)}=\sum_{g=1}^{h}\pi_g\\ \pi_g&=pr(t_{g-1}<t\le t_g),\quad \alpha_h=\sum_{k=1}^{h}\exp(\alpha_k^*+\sum_{s=1}^{q}\delta_s z_{ks}),\quad \alpha_1^*=0\end{aligned} \tag{11.16}$$

for q time varying explanatory variables z_{hs}. Since the exponential cumulative contributions are positive, full monotonicity is guaranteed. Again, random effects for further levels or classifications can be added to the fixed part predictor $X\beta$.

The second line of (11.16) allows time varying covariates to contribute cumulatively to the threshold parameter according to their current values. An alternative is to allow each of the covariates to contribute to the threshold parameter according to its current *mean* value, so that we have

$$\alpha_h = \sum_{k=2}^{h} \exp\left(\alpha_k^* + \sum_{s=1}^{q} \delta_s z_{ks}^*\right), \quad z_{\mathrm{ks}}^* = \frac{1}{k}\sum_{l=1}^{k} z_{ls}, \quad (h > 1) \tag{11.17}$$

and for $h = 1$ we use (11.16). Model (11.16) may be appropriate where covariates exert their effects rapidly, whereas (11.17) may be more appropriate where covariates are slower to affect the timing of the event.

The terms α_k^* in these expressions, together with the intercept β_0, define the baseline hazard function and can be replaced by a smooth function as already mentioned. For ease of interpretation we may wish to centre at least the continuous explanatory variables and we note that an additional intercept among the explanatory variables z, z^*, is not generally required. We have also assumed that for a given explanatory variable the coefficient is invariant over time. If we wish to have dfferent coefficients for different time intervals or groups of intervals this can be accomplished by interacting the dummy variables for the baseline categories with the cexplanatory variables. We can also consider incorporating the first category covariate values in the fixed part of the model and modelling the subsequent covariates as dfferences or changes from these values, so that $\alpha_1 = 0$. This may be preferable in terms of interpreting the parameters. We note that our formulation using a positive monotonic function for the discrete time survival model with time varying covariates corresponds to the traditional modelling of the hazard function using, say, a logistic link function, that has the same monotonicity property. In fact any strictly positive monotonic function can be used instead of the simple exponential. In some situations a bounded function, such as the logistic, may be preferable.

11.12.5 An example using partnership durations

As in Section 11.11.2, the data are based upon partnership histories of female respondents in the National Child Development Study collected retrospectively at ages 33, and also at 42 years. A full description is given in Steele *et al*. (2005). The present analysis uses a subset of the data and explanatory variables, namely the presence of a preschool child, and the age of the respondent. The response event is separation from either cohabitation or marriage and the event time is the partnership duration. The original data are grouped into six month intervals with the maximum number of intervals for a woman being 54. All women with more than 45 intervals are grouped into a final interval to avoid small numbers. Approximately 64 % of the women are still partnered at age 42 and therefore have censored partnership durations.

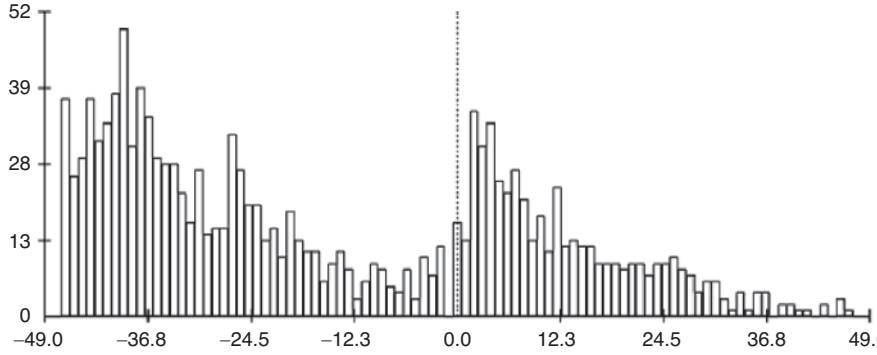


Figure 11.5 Durations of partnerships. negative values are right censored observations.

There are 1000 women with 1331 partnerships. Figure 11.5 shows the distribution of partnership lengths and we see that there are some intervals that only have censored observations. The threshold estimates α_k^* for these then become large and negative and have a negligible contribution to the estimates for α_h. Table 11.6 shows the estimates for fitting both models (11.16) as Model A and (11.17) as Model B. We have omitted the 45 estimates for the thresholds (for details, see Goldstein, 2010a).

In Model B, the overall effect of having a young child at the start of a partnership is to increase the value on the latent normal scale (since it is a covariate belonging to X) and hence to increase the probability that the partnership will end for each given interval, that is decrease the overall probability of remaining in a partnership for all time periods, presumably reflecting the characteristics of a partnership that starts with an existing younger child. It could also reflect unobserved characteristics of women who have children from a previous relationship. The variable, in either case may well be endogenous so that it might be more appropriate to model a fertility process jointly with the partnership dissolution process as in Steele *et al*. (2005),

Table 11.6 MCMC estimates: 500 burn in, 5000 sample. Estimates (SE).

Parameter	Model A	Model B
Intercept	2.39 (0.14)	2.48 (0.16)
Age group 20–24	0.43 (0.11)	0.36 (0.11)
Age group 25–29	0.62 (0.13)	0.53 (0.13)
Age group 30+	0.51 (0.15)	0.40 (0.15)
Young child present (fixed covariate)	−0.71 (0.14)	−0.49 (0.13)
Young child present (time varying)	−0.32 (0.07)	−1.20 (0.27)
Level 2 variance	0.293 (0.093)	0.326 (0.110)
Age group < 20 is the base category		

using a multivariate formulation. Given the presence (or absence) of a young child at the start of the partnership, the effect at separation of the current average of the younger child variable, in effect the proportion of times over the period that there is a younger child present, is to multiply that threshold parameter (additive) contribution (compared to no younger children during the time period) by the mean multiplied by $e^{-1.2} = 0.30$, so that 0.30 is the multiplier when a younger child is always present. This therefore leads to a decrease on the latent normal scale and hence to increase the probability of remaining in a partnership. This suggests that the arrival of a young child during a partnership tends to prolong the partnership as opposed to the effect of starting the partnership with a young child. We also see that the higher the age of starting a partnership the longer the partnership survives. In model A we see that the fixed part contribution for a younger child is greater and the time-dependent effect is larger with a multiplying factor of $e^{-0.32} = 0.73$. The effect is thus to multiply the cumulative base threshold by 0.73 which provides perhaps a more straightforward interpretation than Model B.

The contribution to the threshold value from the time dependent covariate can be written as $\alpha_h = \sum_{k=1}^{h} \delta_h e^{\alpha_k^*}$, $\delta_h = e^{\theta}$ if younger child $\delta_h = 1$ if not, where θ is the coffcient for the explanatory variable. The contribution is therefore a weighted sum of the $e^{\alpha_k^*}$. In the present case, approximately four consecutive intervals (with similar values for $e^{\alpha_k^*}$) with a young child is equivalent to three consecutive intervals without a young child in terms of the contribution to the threshold value.

A further advantage of the nonlinear formulation for the thresholds (11.16) and (11.17) is that where there are high correlations between the x, zvariables the separation into different components can reduce numerical instabilities.

12

Cross-classifie data structures

12.1 Random cross classification

In previous chapters, we have considered only data where the units have a purely hierarchical or nested structure. In many cases, however, a unit may be classified along more than one dimension. An example is students classified both by the school they attend and by the neighbourhood where they live. Figure 12.1 represents this diagrammatically for three schools and four neighbourhoods with between one and six students per school/neighbourhood cell. The cross classification is at level 2 with students at level 1.

Another example is in a repeated measures study where children are measured by different raters at different occasions so that children are essentially cross-classified by raters. If each child has its own set of raters not shared with other children then the cross classification is of occasions and raters within children. This can be represented diagramatically in Figure 12.2 for three children with up to seven measurement occasions and two raters per child. If we adopt the convention that the lowest level at which a response occurs is level 1, then the cross classification here is at level 2 with level three that of the child and this is a level 2 cross classification with only one unit per cell (see also Section 12.5).

If now the same set of raters is involved with all the children the crossing is at the child level, 2, as can be seen in the following diagram with three raters and three children and up to five occasions.

Figure 12.3 is formally the same structure as Figure 12.1 with the level 1 variance being that between occasions.

These basic cross classifications may occur when a simple hierarchical structure breaks down. Consider, for example, a repeated measures design which follows a sample of students over time, and measured once a year, within a set of classes for a single school. We assume first that each class group is taken by the same teacher.

Multilevel Statistical Models: 4th Edition Harvey Goldstein

	School 1	School 2	School 3
Neighbourhood 1	x x x x	x x	x
Neighbourhood 2	x	x x x x x x	x x x
Neighbourhood 3	x x	x	x x x x
Neighbourhood 4	x x x	x x	x x

Figure 12.1 A random cross classificatio at level 2.

The hierarchical structure is then a three level one with occasions grouped within students who are grouped within classes. If we had several schools then schools would constitute the level 4 units. Suppose, however, that students change classes from one year to the next. For three students, three classes and up to three years (occasions) we might have the pattern in Figure 12.4.

Formally this is the same structure as Figure 12.3, that is a cross classification at level 2 for classes by students. We distinguish such a design from the multiple membership designs described in Chapter 13, where mobility can occur at any time and a move is not associated with a new measurement, as here.

Suppose now that, instead of the same teachers taking the classes throughout the study, the classes are taken by a completely new set of teachers every year and where new groupings of students are formed each year too. Such a structure with four different teachers over two years for three students is given in Figure 12.5.

This is now a cross classification of teachers by students at level 2 with occasion as the level 1 unit. We note that most of the cells are empty and that there is at most one level 1 unit per cell so that no independent between occasion variance can

	Child 1	Child 2	Child 3
Occasion:	1 2 3 4 5 6 7	1 2 3 4 5 6 7	1 4 7
Rater 1	x x x x x		
Rater 2	x x x x x		
Rater 3		x x x x x	
Rater 4		x x x x x x	
Rater 5			x x x
Rater 6			x

Figure 12.2 A random cross classificatio at level 2 with one unit per cell.

	Child 1	Child 2	Child 3
Occasion:	1 2 3 4	1 2	1 2 3 4 5
Rater 1	x x x	x	x
Rater 2	x		x x
Rater 3		x	x x

Figure 12.3 A random cross classificatio at level 2.

	Student 1	Student 2	Student 3
Occasion:	1 2 3	1 2	1 2 3
Class/teacher 1	x x	x	x
Class/teacher 2	x		
Class/teacher 3		x	x x

Figure 12.4 Students changing classes/teachers.

be estimated. Raudenbush (1993) gives an example of such a design, and provides details of an EM estimation procedure for 2-level 2-way cross classifications with worked examples.

We can have a design which is a mixture of those given by Figure 12.4 and Figure 12.5 where some teachers are retained and some are new at each occasion. In this case we would have a cross classification of teachers by students at level 2 where some of the teachers only had observations at one occasion. More generally, we can have an unbalanced design where each teacher is present at a variable number of occasions.

With two occasions where we have the same teachers or intact groups we can formulate an alternative cross-classification design which may be more appropriate in some cases. Instead of cross classifying students by teachers, we consider cross classifying the set of all teachers at the first occasion by the same set at the second occasion. Consider Figure 12.6, where we have 22 students who are nested within the cross classification of teachers at each occasion. The difference between this design and that in Figure 12.4 is analogous to the difference between a two-occasion longitudinal design where a second occasion measurement is regressed on a first occasion measurement and the two-occasion repeated measures design where a measurement

		Student 1	Student 2	Student 3
Year:		1 2	1 2	1 2
Teacher 1	1	x	x	
Teacher 2				x
Teacher 3	2	x		x
Teacher 4			x	

Figure 12.5 Students changing teachers and groups.

		Occasion 2 Teacher 1	Teacher 2	Teacher 3
Occasion 1	Teacher 1	x x x x x	x	x x
	Teacher 2	x x	x x x x	
	Teacher 3	x	x x x	x x x x

Figure 12.6 Teachers cross classifie by themselves at two occasions with responses from 22 students.

is related to age or time. In Figure 12.6, we are concerned with the contribution from each occasion to the variation in, say, a test score measured at occasion 2, and we might consider any first occasion meaurement as an explanatory variable. In Figure 12.4, on the other hand, although we could fit a separate between-teacher variance for each occasion, the response variable is essentially the same one measured at each occasion.

12.2 A basic cross-classifie model

Goldstein (1987a) sets out the general structure of a model with both hierarchical and cross classified structures. We consider first the simple model of Figure 12.1 with variance components at level 2 and a single variance term at level 1.

We shall refer to the two classifications at level 2 using the subscripts j_1, j_2 and, in general, parentheses will group classifications at the same level. We write the model as

$$y_{i(j_1 j_2)} = X_{i(j_1 j_2)}\beta + u_{1 j_1} + u_{2 j_2} + e_{i(j_1 j_2)} \tag{12.1}$$

The covariance structure at level 2 can be written in the following form

$$\begin{aligned} &\operatorname{cov}(y_{i(j_1 j_2)} y_{i'(j_1 j_2')}) = \sigma^2_{u_1} \\ &\operatorname{cov}(y_{i(j_1 j_2)} y_{i'(j_1' j_2)}) = \sigma^2_{u_2} \\ &\operatorname{var}(y_{i(j_1 j_2)}) = \operatorname{cov}(y_{i(j_1 j_2)} y_{i'(j_1 j_2)}) = \sigma^2_{u_1} + \sigma^2_{u_2} \end{aligned} \tag{12.2}$$

We note that if there is no more than one level 1 unit per cell model (12.1) is still valid.

The level 2 variance is the sum of the separate classification variances, the covariance for two level 1 units in the same classification is equal to the variance for that classification and the covariance for two level 1 units which do not share either classification is zero. If we have a model where random coefficients are included for either or both classifications, then analogous structures are obtained. We can also add further ways of classification with obvious extensions to the covariance structure. The essence of (12.2) is that the variances from each classification are additive. We may, however, have more complex models where the total level 2 variance is more complex and cannot be described by the sum of separate variances. Such interactive models can be fitted within the framework set out in Appendix 12.1, but some of the simplicity of interpretation achieved by assuming (12.2) will then be lost. For example, we can fit separate random effects, with an associated variance, for a subset of cells, and in the extreme case we may fit a separate (undifferentiated) random effect for each cell with a single variance term. We return to this below.

Appendix 12.1 shows how cross-classified models can be specified and estimated efficiently, based upon a hierarchical formulation using the IGLS algorithm. For large problems, however, where two or more cross classifications contain large numbers of

units, this procedure becomes unwieldy and in a later section we show how MCMC estimation can handle such situations and introduce a more general notation.

12.3 Examination results for a cross classificatio of schools

The data consist of scores on school leaving examinations obtained by 3435 students who attended 19 secondary schools cross-classified by 148 primary schools in Fife, Scotland (Paterson, 1991). Before their transfer to secondary school at the age of 12 each student obtained a score on a verbal reasoning test, measured about the population mean of 100 and with a population standard deviation of 15.

The model is as follows:

$$\begin{aligned} &y_{i(j_1 j_2)} = \beta_0 + \beta_1 x_{1i(j_1 j_2)} + u_{1j_1} + u_{2j_2} + e_{i(j_1 j_2)} \\ &u_{1j_1} \sim N(0, \sigma^2_{u_1}), \quad u_{2j_2} \sim N(0, \sigma^2_{u_2}), \quad e_{i(j_1 j_2)} \sim N(0, \sigma^2_e) \end{aligned} \tag{12.3}$$

and the results are given in Table 12.1. Random coefficients for verbal reasoning were also fitted but the associated variances are estimated as zero.

Ignoring the verbal reasoning score (Model A), we see that the between-primary school variance is estimated to be more than three times that between secondary schools. One reason for this may be that the secondary schools are on average larger and more homogeneous socially than primary schools, so that the variance is smaller. To study further the issue of school size and composition we could make the between-school variance a function of such variables, but these are not available in the present dataset.

When the verbal reasoning score (Model B) is added to the fixed part of the model the between secondary school variance becomes very small, the between primary school variance is also considerably reduced and the level 1 variance also. The third analysis (Model C) shows the effect of removing the cross classification by primary

Table 12.1 Analysis of Examination Scores by Secondary and Primary school attended. The subscript 1 refers to primary and 2 to secondary school.

Parameter	Estimate (s.e.)	Estimate (s.e.)	Estimate (s.e.)
Fixed	Model A	Model B	Model C
Intercept	5.50	5.98	5.99
Verbal Reasoning	–	0.16 (0.003)	0.16 (0.003)
Random			
$\sigma^2_{u_1}$	1.12 (0.20)	0.27 (0.06)	–
$\sigma^2_{u_2}$	0.35 (0.16)	0.011 (0.021)	0.28 (0.06)
σ^2_e	8.1 (0.2)	4.25 (0.10)	4.26 (0.10)

school. The between secondary school variance is now only a little smaller than in analysis A without verbal reasoning score. Using analysis C alone, which is typically the case with school effectiveness studies which control for initial achievement, we would conclude that there were important differences between the progress made in secondary schools. From analysis B, however, we see that most of this is explained by the primary schools attended. Of course, the verbal reasoning score is only one measure of initial achievement, but these results illustrate that in such educational data, adjusting for achievement at a single previous time may not be adequate.

12.4 Interactions in cross classification

Consider the following extension of (12.1):

$$y_{i(j_1 j_2)} = X_{i(j_1 j_2)}\beta + u_{1 j_1} + u_{2 j_2} + u_{3(j_1 j_2)} + e_{i(j_1 j_2)}. \tag{12.4}$$

We have now added a general 'interaction' term to the model which was previously an additive one for the two random effects. The usual specification for such a random interaction term is that it has a simple variance $\sigma^2_{u_{(12)}}$ across all the level 2 cells (Searle *et al*, 1992). The adequacy of such a model can be tested against an additive model using a likelihood ratio test criterion. For the example, in Table 12.1 this interaction term is estimated as zero. While this indicates that the cross classification is adequate, because the between-secondary-school variance is so small, we would not expect to be able to detect such an interaction.

Extensions to this model are possible by adding random coefficients for the interaction component, just as random coefficients can be added to the additive components. For example, the gender difference between students may vary across both primary and secondary schools in the example given in Section 12.3 and we can fit an extra variance and covariance term for this to both the additive effects and the interaction.

12.5 Cross classification with one unit per cell

Some interesting models occur when there is only one level 1 unit per cell of a level 2 cross classification. This should be distinguished from the case where a level 2 cross classification happens to produce no more than 1 level 1 unit in a cell as a result of sampling. Thus, (12.4), where there are some cells with more than one level 1 unit, allows us to obtain separate estimates for the level 1 variance and the level 2 interaction. If there is only one level 1 unit per cell then this interaction is confounded with the level 1 variance.

So called 'generalisability theory' models (Cronbach and Webb, 1975) can be formulated as a cross classification with one level 1 unit per cell. The basic model is one where a test or other instrument consisting of a set of items, for example ratings or questions, is administered to a sample of individuals. The individuals are therefore cross-classified by the items and may be further nested within schools etc. at higher levels. In educational test settings the item responses are often binary, so that we

would apply the methods of Chapter 4 to the present procedures in a straightforward way. Since each individual can only respond once to each item, this is an example where we cannot directly detect a level 2 interaction.

12.6 Multivariate cross-classifie models

For multivariate models the responses may have different structures. Thus, in a bivariate model one response may have a 2-level hierarchical structure and the other may have a cross classification at level 2. Suppose we measure the height and the mathematics attainment of a sample of students, at several occasions, from a sample of schools. The mathematics attainment is assessed by a different set of teachers in each school and the heights are measured by a single anthropometrist. For the mathematics scores there is a cross classification of students by teachers within each school, whereas for height there is a 2-level hierarchy with students nested within schools. Height and mathematics attainment will be correlated at both the student and the school level. We can write a model for this structure as follows:

$$
\begin{aligned}
&y_{h(i_1i_2)j} = \delta_{1h}(X_{1(i_1i_2)j}\beta_1 + u_{1j} + e_{1i_1j} + e_{1i_2j}) + \delta_{2h}(X_{2i_1j}\beta_2 + u_{2j} + e_{2i_1j}) \\
&\operatorname{cov}(u_{1j}u_{2j}) = \sigma_{u12} \quad \operatorname{cov}(e_{1i_1j}e_{2i_1j}) = \sigma_{e12} \\
&\delta_{1h} = \textit{1 if mathematics},\ \textit{0 if height},\ \delta_{2h} = 1 - \delta_{1h}
\end{aligned}
\tag{12.5}
$$

where all other covariances are zero.

We have already mentioned that cross-classified models can have a discrete response. We can also fit, for example, time series models, as discussed in Chapter 5, and in general cross-classified structures can incorporate all the types of models which can be fitted for purely hierarchical structures.

12.7 A general notation for cross classificatio

In Figure 12.7, we set out a simple 'classification diagram' as introduced by Browne *et al.* (2001). It allows us to classify data structures as hierarchical or crossed or combinations of these at different levels. Boxes represent unit classifiers and those at the same horizontal level in a cross classification are at the same level conceptually. In terms of the model, we also have a simplified notation; for a basic variance components model we write

$$
\begin{aligned}
&y_i^{(1)} = (X\beta)_i + u^{(2)}_{school(i)} + u^{(1)}_{student(i)} \quad school(i) \in (1, \ldots, J) \\
&student(i) \in (1, \ldots, N) \\
&u^{(2)}_{school(i)} \sim N(0, \sigma^2_{u(2)})) \qquad u^{(1)}_{student(i)} \sim N(0, \sigma^2_{u(1)}) \quad i = 1, \ldots, N
\end{aligned}
\tag{12.6}
$$

In (12.6) we now have two *classification* , with students as classification 1 and schools as classification 2. The subscript i is attached to the lowest level units and uniquely identifies every measurement and random effect. Thus *school(i)* is the school that

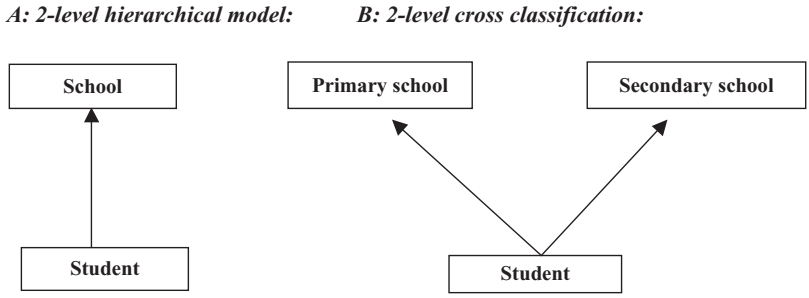


Figure 12.7 Classificatio diagrams for hierarchical and cross classifie structures.

student i belongs to and we refer to *school(i)* as a classification function that maps the lowest level units, students, onto schools. *Student(i)* is the identity function that maps students onto themselves. The superscript denotes the classification, where the lowest is numbered 1 etc. Together with Diagram A in Figure 12.7, Equation (12.6) completely specifies the model structure.

We can now rewrite (12.1) as

$$\begin{aligned} &y_i^{(1)} = (X\beta)_i + u^{(3)}_{neighbourhood(i)} + u^{(2)}_{school(i)} + u^{(1)}_{student(i)} \\ &neighbourhood(i) \in (1, \ldots, J_3) \quad school(i) \in (1, \ldots, J_2) \\ &student(i) \in (1, \ldots, N)) \qquad u^{(3)}_{neighbourhood(i)} \sim N(0, \sigma^2_{u(3)}) \\ &u^{(2)}_{school(i)} \sim N(0, \sigma^2_{u(2)}) \quad u^{(1)}_{student(i)} \sim N(0, \sigma^2_{u(1)}) \quad i = 1, \ldots, N \end{aligned} \tag{12.7}$$

which together with diagram B in Figure 12.7 completely specifies the cross classified model. Since the only subscript is that for the lowest level units the notation can be extended indefinitely for any number of crossed or hierarchical classifications.The earlier notation as in (12.1) becomes very cumbersome when many classifications exist. Browne *et al*. (2001) give a comprehensive treatment of this general notation which, as we shall see in Chapter 13, extends readily to handle multiple membership data structures. These authors also adopt the convention of dropping the (1) superscript for level 1 responses and effects. This provides somewhat greater clarity for models such as (12.7) but there are models, for example with multivariate responses at several levels, where it is useful to retain the level 1 superscript.

12.8 MCMC estimation in cross-classifie models

The extension of MCMC methods to cross classifications is straightforward since we simply introduce further steps for each extra classification that samples residuals and covariance matrices from the relevant posterior distribution. Thus, for example, for (12.9), the Gibbs algorithm in Section 2.13 becomes

Step 1 Sample a new set of fixed effects (β).

Step 2 Sample a new set of neighbourhood residuals $\{u^{(3)}\}$.

Step 3 Sample a new set of school residuals $\{u^{(2)}\}$.

Step 4 Sample a new neighbourhood classification variance.

Step 5 Sample a new school classification variance

Step 6 Sample a new level 1 variance.

Step 7 Compute the level 1 residuals by subtraction.

Similar modifications to MCMC algorithms allow cross classified models with discrete responses etc. For starting values typically it will be possible to use estimates obtained from fitting a purely hierarchical model for each single higher level classification in turn. If we use MCMC with default Gamma priors to fit the Fife examination data we obtain results very similar to those in Table 12.1.

Appendix 12.1 IGLS estimation for cross-classifie data

12.1.1 An efficien IGLS algorithm

We illustrate the procedure using a 2-level model with crossing at level 2.

The 2-level cross-classified model, using the notation in Appendix 2.1, can be written as

$$y_{i(j_1 j_2)} = X_{i(j_1 j_2)}\beta + \sum_{h=1}^{q_1} z_{1hij_1} u_{1hj_1} + \sum_{h=1}^{q_2} z_{2hij_2} u_{2hj_2} + e_{i(j_1 j_2)} \tag{12.1.1}$$

Parentheses group the ways of classification at each level. We have two sets of explanatory variables, type 1 and type 2, for the random components defined by the columns of $Z_1(n \times p_1 q_1)$, $Z_2(n \times p_2 q_2)$ where p_1, p_2 are respectively the number of categories of each classification.

$$\begin{array}{llll} Z_1 = \{z_{1hij_1}\}, & Z_2 = \{z_{2hij_2}\} & & \\ z_{1hij_1} = z_{1him} & if \quad j_1 = m, & for\ m-th\ type\ 1\ level\ 2\ unit, & 0\ otherwise \\ z_{2hij_2} = z_{2him} & if \quad j_2 = m, & for\ m-th\ type\ 2\ level\ 2\ unit, & 0\ otherwise \end{array}$$

These variables are dummy variables where for each level 2 unit of type 1 we have q_1 random coefficients with covariance matrix $\Omega_{(1)2}$ and likewise for the type 2 units. To simplify the exposition we restrict ourselves to the variance component case where we have

$$\begin{aligned} &\Omega_{(1)2} = \sigma^2_{(1)2}, \quad \Omega_{(2)2} = \sigma^2_{(2)2} \\ &E(\tilde{Y}\tilde{Y}^T) = V_1 + Z_1(\sigma^2_{(1)2} I_{(p_1)}) Z_1^T + Z_2(\sigma^2_{(2)2} I_{(p_2)}) Z_2^T \end{aligned} \tag{12.1.2}$$

Consider Figure 12.1, where we sort the data by school which is classification 1. The explanatory variables will have the structure in Figure 12.1.1 for the first eight students.

It is clear that the second term in (12.1.2) can be written as

$$Z_1(\sigma^2_{(1)2} I_{(p_1)}) Z_1^T = J\sigma^2_{(1)2} J^T$$

where J is a (n $\times$ 1) vector of ones. The third term is of the general form $Z_3 \Omega_3 Z_3^T$, namely a level 3 contribution where in this case there is only a single level 3 unit and with no covariances between the random coefficients of the Z_{2h} and with the variance terms constrained to be equal to a single value, $\sigma^2_{(2)2}$.

More generally, we can specify a level 2 cross classified variance components model by modelling one of the classifications as a standard hierarchical component and the second as a set of dummy explanatory variables, one for each category, with the random coefficients uncorrelated and with variances constrained to be equal.

i,j_1	i,j_2	Z_{11}	Z_{12}	Z_{13}	Z_{14}	Z_{21}	Z_{22}	Z_{23}
1,1	1,1	1	0	0	0	1	0	0
2,1	2,1	1	0	0	0	1	0	0
3,1	3,1	1	0	0	0	1	0	0
4,1	4,1	1	0	0	0	1	0	0
5,1	1,2	1	0	0	0	0	1	0
6,1	2,2	1	0	0	0	0	1	0
7,1	1,3	1	0	0	0	0	0	1
1,2	2,1	0	1	0	0	1	0	0

Figure 12.1.1 Explanatory variables for level 2 cross-classificatio of Figure 12.1.

If this second (type 2) classification has further explanatory variables with random coefficients as in (12.1.1) then we form extended dummy variable 'interactions' as the product of the basic dummy variables and the further explanatory variables with random coefficients, so that these coefficients have variances and covariances within the same type 2 level 2 unit but not across units. In addition the corresponding variances and covariances are constrained to be equal.

To extend this to further ways of classification we add levels. Thus, for a three-way cross classification at level 2 we choose one classification, typically that with the largest number of categories, to model in standard hierarchical fashion at level 2, the second to model with coefficients random at level 3 as above and the third to model in a similar fashion with coefficients random at level 4. The same principle applies to cross classifications at level 1 nested within level 2 units. The level 1 cross classification is modelled as a 2-level hierarchy with the original level 2 units becoming level 3 units. We can also allow simultaneous crossing at more than one level. Thus for example, if there is a 2-way cross classification at level 1 and a 3-way cross classification at level 2, we will require five levels, the first two describing the level 1 cross classification and the next three describing the level 2 cross classification. Further details are given by Rasbash and Goldstein (1994).

12.1.2 Computational considerations

Analysis A in Table 12.1 took about ten times longer than analysis C. This relative slowness is due to the size of the single level 3 unit which contains all the 3435 level 1 units. For very much larger problems the computing considerations will become of greater concern, so that some procedure for speeding up the computations is needed.

In the present analysis there are 120 cells of the cross classification which contain only one student. If we eliminate these data from the analysis, we obtain two disjoint subsets containing 14 and five secondary schools. There are a further 24 cells containing two students and if these are removed we obtain six disjoint subsets the largest of which contains eight secondary schools. Table 12.1.1 shows the estimates from the resulting analyses.

Table 12.1.1 Examination scores for Secondary by Primary school classification omitting small cells.

Parameter	Estimate (s.e.)	Estimate (s.e.)
Fixed	≤ 1 student	≤ 2 students
Intercept	6.00	6.00
Verbal reasoning	0.16 (0.003)	0.16 (0.003)
Random		
$\sigma^2_{u(1)0}$	0.27 (0.06)	0.25 (0.06)
$\sigma^2_{u(2)0}$	0.004 (0.021)	0.028 (0.030)
σ^2_e	4.28 (0.11)	4.29 (0.11)

The only substantial difference is in the between secondary school variance which is anyway poorly estimated. In many cases such separations may not be possible and MCMC estimation (Section 12.8) becomes relatively fast and is the preferred method. It also has the advantage that it is uneccessary to eliminate any data as has been done here.

13

Multiple membership models

13.1 Multiple membership structures

In some circumstances, units can be members of more than one higher level unit. An example is friendship patterns where at any time individuals can be members of more than one friendship group. In an educational system, students may attend more than one institution. In all such cases we shall assume that for each higher level unit to which a lower level unit belongs there is a known weight (summing to 1.0 for each lower level unit), which represents, for example, the amount of time spent in that unit. The choice of weights may be important, and so we might carry out a sensitivity analysis to determine how alternative choices of weights affect inferences, and using MCMC estimation we can use the DIC to compare models with different sets of weights. Appendix 13.1 sets out the MCMC sampling steps.

We may also have data where there is some uncertainty about which higher level unit some lower level units belong to; for example, in a survey of students, information about their neighbourhood of residence may only be available for a few students for larger geographical units. For these cases, it may be possible to assign a weight for each of the constituent neighbourhoods which is in effect a probability of belonging to each based upon available information. We will refer to such a structure as a missing identification structure for which the multiple membership estimation techniques are readily adapted (see Section 13.7).

Consider first the 2-level variance components model with each level 1 unit belonging to at most two level 2 units, where the j_1, j_2 subscripts now refer to the same type of unit, so that we do not require separate subscripts (1,2,3...) for the u_j terms.

$$\begin{aligned} &y_{i(j_1 j_2)} = X_{i(j_1 j_2)}\beta + w_{1ij_1}u_{j_1} + w_{2ij_2}u_{j_2} + e_{i(j_1 j_2)} \\ &w_{1ij_1} + w_{2ij_2} = 1 \end{aligned} \tag{13.1}$$

Multilevel Statistical Models: 4th Edition Harvey Goldstein

The overall contribution at level 2 is the weighted sum over the level 2 units to which each level 1 unit belongs. This leads to the following covariance structure

$$\begin{aligned}
&\text{var}(y_{i(j_1 j_2)}) = (w^2_{1ij_1} + w^2_{2ij_2})\sigma^2_u + \sigma^2_e \\
&\text{cov}(y_{i(j_1 j_2)} y_{i'(j_1 j_2)}) = (w_{1ij_1} w_{1i'j_1} + w_{2ij_2} w_{2i'j_2})\sigma^2_u \\
&\text{cov}(y_{i(j_1 j_2)} y_{i'(j'_1 j_2)}) = w_{2ij_2} w_{2i'j_2} \sigma^2_u
\end{aligned}$$

For individuals belonging to more than one unit, the contribution to the total variation of the higher level units is less. For example, suppose there are just two higher level units.

$$\text{If} \quad w_{i1} = w_{i2} = 0.5, \quad \text{var}\left(\sum_h w_{ih} u_h\right) = \sigma^2_u / 2 \tag{13.2}$$

In other words, what we see is an average over several schools that will naturally have less variation than that associated with attendance at a single school. This has important implications for models which ignore large numbers of multiple memberships. Suppose in a study of school examination results, a substantial number of students have spent time in several schools but they are assigned only to the final school where they took the examination. (This is discussed with examples by Goldstein *et al*., 2007b.) The observed level 2 variation will be less than the true level 2 variation for these students, as demonstrated in (13.2). The effect will be to underestimate the true level 2 variation and thus to produce biased estimates for school level residuals. It would seem that almost all so-called 'school effectiveness' studies, including those used for illustration in earlier chapters are subject to such biases, although it is not always clear how important these are.

Where we have level 2 variables in the fixed part of the model, we would also usually wish to apply the weights to these so that the predictor reflects the weights attached to the level 2 unit. For example, as students move between schools and we have a predictor such as the size of school that the student is in, then we may wish to construct a weighted average of these.

13.2 Notation and classification for multiple membership structures

Using the general notation introduced for cross classifications given in Chapter 12, we can write (13.1) as

$$\begin{aligned}
&y_i = (X\beta)_i + \sum_{j \in school(i)} w^{(2)}_{i,j} u^{(2)}_j + e_i \\
&u^{(2)}_j \sim N(0, \sigma^2_{u(2)}), \quad e_i \sim N(0, \sigma^2_e) \\
&school(i) \in \{1, \ldots J\}, \quad i = 1, \ldots N
\end{aligned} \tag{13.3}$$

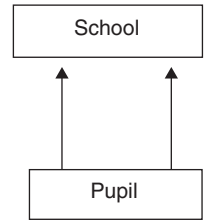


Figure 13.1 Classificatio diagram for 2-level multiple membership structure.

where we have omitted the superscript (1) for the level 1 classification for convenience. We note that now *school(i)* refers to all the schools that unit i belongs to. The classification diagram for (13.3) is given in Figure 13.1.

The double arrow indicates a multiple membership relationship. For likelihood-based estimation we set up a set of J variables where the value is $w_{i,j}^{(2)}$ for the j-th unit with a nonzero weight for level 1 unit i, and zero otherwise. As in the cross-classified case, these then have random coefficients with variances constrained to be equal (Appendix 12.1). For MCMC estimation, the weights are incorporated into the sampling of residuals at each iteration. Details are given in Browne *et al*. (2001a).

We now illustrate the analysis of a multiple membership model with an example of veterinary data.

13.3 An example of salmonella infection

The data concern outbreaks of salmonella infection in Danish chicken flocks between 1995 and 1997. The response is whether or not salmonella is present in a (slaughtered) 'child' flock. Each flock is kept in a house within a farm, the hierarchical component, and is created from a mixture of up to six parent flocks, the multiple membership component. There is also the cross classification of the parent flocks by house within farms. In addition, there are four chicken hatcheries for which we fit three dummy variables, together with two dummy variables for year. There are 200 parent flocks, 304 farms, 725 houses and 10,127 child flocks of chickens. The classification diagram is given in Figure 13.2 and the model is

$$\begin{aligned} &y_i \sim bin(1, \pi_i) \\ &\text{logit}(\pi_i) = (X\beta)_i = u^{(2)}_{house(i)} + u^{(3)}_{farm(i)} + \sum_{j \in flo\ k(i)} w^{(4)}_{i,j} u^{(4)}_j \\ &u^{(2)}_{house(i)} \sim N(0, \sigma^2_{u(2)}),\ u^{(3)}_{farm(i)} \sim N(0, \sigma^2_{u(3)}),\ u^{(4)}_j \sim N(0, \sigma^2_{u(4)}) \end{aligned} \tag{13.4}$$

The results of fitting (13.4) using MCMC are given in Table 13.1. The year 1995 and hatchery 1 are taken as base categories. Default inverse Gamma priors are used for variances and MH estimation is based upon 50,000 iterations with a burn in of 20,000.

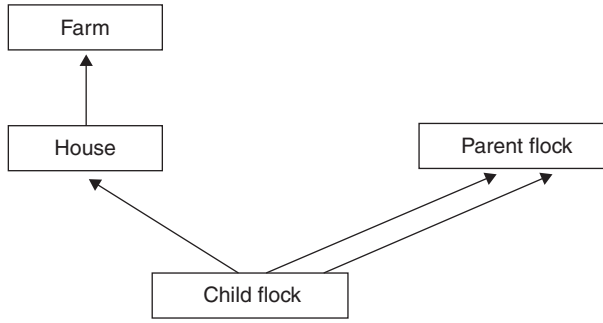


Figure 13.2 Classificatio diagram for salmonella data.

Table 13.1 MCMC estimates for salmonella data.

Fixed	Estimate (s.e.)
Intercept	−2.33 (0.22)
1996–1995	−1.24 (0.17)
1997–1995	−1.16 (0.19)
Hatchery 2 – Hatchery 1	−1.73 (0.26)
Hatchery 3 – Hatchery 1	−0.20 (0.25)
Hatchery 4 – Hatchery 1	−1.06 (0.38)
Variances	
Parent flock	0.88 (0.18)
Farm	0.92 (0.20)
House	0.20 (0.11)

It is clear that most of the variation is between farms and between parent flocks, with some large hatchery differences. Residuals can be estimated for all effects so that extreme farms and parent flocks can be identified. Further details can be found in Browne *et al*. (2001a).

13.4 A repeated measures multiple membership model

Goldstein *et al*. (2000a) consider repeated measures household data where individuals move from household to household over a period of five years divided into six-monthly occasions. The response is the average length of time each individual has spent in all households since the start of the study, up to the current occasion. Thus, for example after three years, an individual who has spent a year in the first household and two years in the next will have a response value of 1.5 years. A household is not a constant unit but becomes a new one when any individual

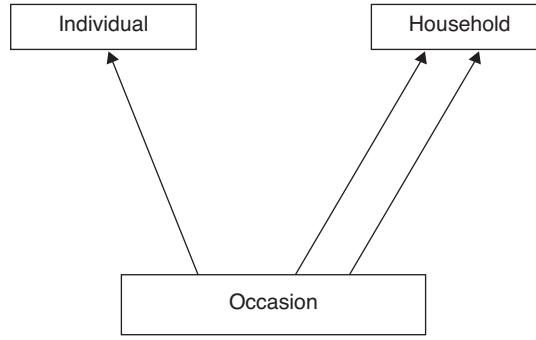


Figure 13.3 A repeated measures multiple membership model.

leaves or enters. The classification diagram is given in Figure 13.3 and the model in (13.5). The lowest level unit is the measurement occasion indexed by i, and these are nested within individuals. At each occasion the measurement has a contribution from a set of households so that occasions are formally multiple members of households.

$$\begin{aligned} y_i &= (X\beta)_i + u^{(2)}_{individual(i)} + \sum_{j \in household(i)} w^{(3)}_{i,j} u^3_j + e_i \\ u^{(2)}_{individual(i)} &\sim N(0, \sigma^2_{u(2)}),\ u^3_j \sim N(0, \sigma^2_{u(3)}),\ e_i \sim N(0, \sigma^2_e) \end{aligned} \tag{13.5}$$

The analysis given in Goldstein *et al*. (2000a) shows a very large between household variance compared to the between-individual variance, together with effects for age and nationality.

13.5 Individuals as higher level units

In all our examples of cross classifications and multiple membership structures so far considered, individuals, or occasions in repeated measures models, are the lowest level units. In some structures, however, they can become higher level units with traditional higher level units becoming level 1 units. Consider the following educational structure.

Suppose we have students who are assessed individually and also in groups, where the measurement is made at the group level, and we wish to know what individuals contribute to the group response – in particular, the correlation between the individual assessment and this contribution. Suppose also that we have data where each individual in a sample is assessed, over time, in several groups with differing

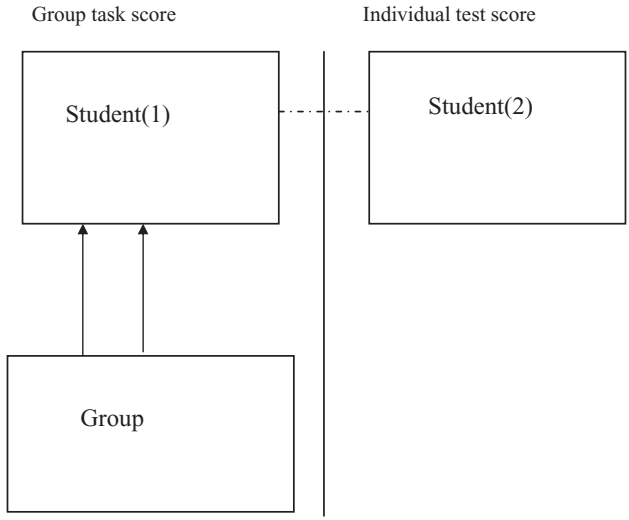


Figure 13.4 Classificatio diagram with two responses.

compositions. We have two responses and can write the following model:

$$
\begin{aligned}
y_{1i}^{(1)} &= (X_1\beta_1)_i + \sum_{j \in individual(i)} w_{i,j}^{(2)} u_{1j}^{(2)} + e_i^{(1)} \text{ (group level response)} \\
y_{2j}^{(2)} &= (X_2\beta_2)_j + u_{2j}^{(2)} \text{ (individual level response)} \\
\begin{pmatrix} u_{1j}^{(2)} \\ u_{2j}^{(2)} \end{pmatrix} &\sim N(0, \Omega_u), \Omega_u = \begin{pmatrix} \sigma_{u1(2)}^2 & \\ \sigma_{u12(2)} & \sigma_{u2(2)}^2 \end{pmatrix}, e_i^{(1)} \sim N(0, \sigma_{e(1)}^2)
\end{aligned}
\tag{13.6}
$$

Level 1 is now the group level so that groups are considered as multiple members of individuals and it follows for identification purposes that we require more groups than individuals. We shall take up this issue again later. The classification diagram for (13.6) is given in Figure 13.4.

The dotted line linking the two responses is used to indicate a multivariate model, with the vertical line separating the two parts of the model.

This model allows us to estimate the individual level residuals $u_{1j}^{(2)}$, contributing to the group level response, and also the correlation between these and the individual level residuals associated with the individual level response $u_{2j}^{(2)}$. The model can be extended straightforwardly to incorporate random coefficients and discrete responses. Other possible applications of such models are in applications where teams with varying compositions of individuals are formed in sporting activities or for work tasks. We now look at an application of this model to assessing the contribution that individuals and departments in universities make to research grant award grades.

13.5.1 Example of research grant awards

The following analyses are based upon research grant applications submitted to the UK Economic and Social Research Council in the period 2001–07. Each application is awarded a grade whether or not it is finally funded and this grade has been converted to a numerical score. (Full details can be found in Goldstein, 2010b.) The response is the awarded grade score so that the application is the level 1 unit. The overall contribution of the applicants is the weighted sum of the applicant effects u_j with the weights summing to 1.0 for each application. Thus the applicant, formally, is the level 2 unit. Fitting this model allows properly for the grouping of applications within applicant and we can compute the applicant effects, together with confidence intervals, as required.

In fact, we cannot directly fit this model to our data because we have more applicants (level 2 units) than applications (level 1 units) so that we are not able to identify individual applicant effects. Instead, therefore, we choose the university department as level 2 and for each department, for each application we treat it as a level 2 unit with a weight attached that is the sum of the weights of the applicants belonging to that department. Such a formulation is not entirely satisfactory since, in this model, there are applicants who will appear in more than one application, whereas the model assumes effectively that the applicants are independently 'sampled' from each department. To take account of this applicant effect, however, would involve fitting a term for applicant in the model leading to a cross classification of applicant and application, but as we have seen the data do not allow us to fit a random effect classification for applicant. Furthermore, since applicants move between departments this would constitute a cross classification of applicant by department also. In general a failure to account for such non-independence will produce standard error estimates that are downwardly biased and we need to be aware of this in our interpretations.

We can write our model as follows:

$$\begin{aligned} y_i &= \beta_0 + \sum_k \beta_k \delta_k + \sum_{j \in department(i)} w_{i,j} u_j^{(2)} + e_i \\ u_j^{(2)} &\sim N(0, \sigma^2_{u(2)}),\ e_i \sim N(0, \sigma_e^2) \end{aligned} \tag{13.7}$$

where j indexes department and $\{\delta_k\}$ are six dummy variables for the set of seven disciplines to which departments have been allocated. The final total number of separate department codes is 1698 with a total of 2698 applications. Table 13.2 presents results based on only those disciplines with at least 100 applications.

Figure 13.5 shows the 100 smallest and 100 largest department residual estimates with conventional 95 % confidence intervals and it is clear that little separation is possible.

13.6 Spatial models

In spatial modelling, measurements on individuals within an area are assumed to depend both on that area's effect and the effects of surrounding areas. Thus, suppose

Table 13.2 2-level multiple membership model including major disciplines with numerical final grade as response. MCMC estimates. Burn in = 500, chain length = 5000.

Parameter	Estimate	Standard error
Intercept (Base category Economics)	7.17	0.11
Management	−0.69	0.17
Social Policy	−0.54	0.20
Education	−0.24	0.11
Sociology	0.06	0.15
Human Geography	0.10	0.18
Psychology	0.16	0.13
Level 2 variance (Department)	0.45	0.11
Level 1 variance (Application)	2.37	0.09

we are studying health status, we might start by assuming a 2-level model with an individual random effect and a random effect from the area in which a person lives. It may, however, be unrealistic to ignore the effects of surrounding areas (neighbours) on health status and to include these we might have a model such as the following

$$\begin{aligned} y_i &= (X\beta)_i + \sum_{j \in neighbourhs(i)} (w_{i,j}^{(2)} u_j^{(2)}) + u_{area(i)}^{(3)} + e_i \\ u_j^{(2)} &\sim N(0, \sigma_{u(2)}^2),\ u_j^{(3)} \sim N(0, \sigma_{u(3)}^2),\ e_i \sim N(0, \sigma_e^2) \end{aligned} \tag{13.8}$$

We are here modelling separate random effects for the area of residence (classification 3) and the neighbouring areas (classification 2). In some cases all the areas

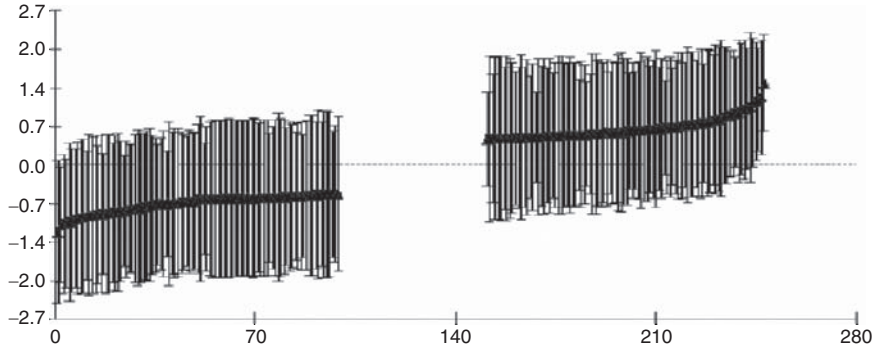


Figure 13.5 One hundred lowest and highest ranked residuals for multiple membership model using all departments, with 95% confidenc intervals.

in a sample are neighbours to all other areas so that for each area there are two random effects in a variance components model and we can fit a covariance between them. Langford *et al*. (1999) do this for some mortality data in Scotland and show a high correlation suggesting that areas have similar relative effects when they are contributing as neighbours and as the area of residence. In this particular case one might replace the two classifications by a single one with one random effect for each area. Leyland (2001) gives further examples and extends the model to consider a multivariate response where covariances between different responses are modelled for both area of residence and for neighbours. Browne *et al*. (2001a) compare the multiple membership spatial model with the traditional conditional autoregressive (CAR) model for an example with Poisson responses for disease prevalence data.

We can extend (13.8) for different response types and with further random coefficients. There are implicit weights in (13.8); equal weights attached to the neighbours, and also an equal weight for the total neighbourhood contribution and the residence area contribution. While the weights attached to neighbours are often chosen to be equal, other choices based upon distance are also possible. It is also possible to give more, or less, weight to the area of residence compared to the neighbouring areas and such choices can be compared using, for example the DIC statistic.

A further extension of these models is to consider as a further classification, say, the area where an individual works: if an individual works in more than one area then this will be another multiple membership classification.

These spatial models can be used to model interacting institutions. Thus, for example, the effects of schooling may come not only from the school that a student belongs to but also from neighbouring schools that may be competing for resources etc. This would lead to a model such as (13.8) where each school will affect student outcomes for its own students and also for those in neighbouring schools. One of the aims of such a model might be to try to discover if there were characteristics of schools that could explain 'neighbouring' effects. In addition, of course, we can jointly model the effects of other classifications such as area of residence. (In Chapter 17, we look at other ways in which relationships across level 2 units can be modelled.)

13.7 Missing identificatio models

A common problem in multilevel data is where the identification of a unit, usually at a higher level, is uncertain. Thus, for example we may have information from a household survey where, for some households, the area is unknown because the information has been lost or withheld, perhaps for confidentiality reasons. In a longitudinal study of schooling we may know that a student has changed school but not which school she came from. In such cases, if we can attach a *probability* of unit membership for each available higher level unit, then we can use these probabilities in our model. If we denote these probabilities by $\pi_{i,j}$, then for a variance components model with normally distributed residuals we can write

$$
\begin{aligned}
& y_i = (X\beta)_i + \sum_{j \in unit(i)} u_j^{(2)} \sqrt{\pi_{i,j}^{(2)}} + e_i \\
& \text{var}\left(\sum_{j \in unit(i)} u_j^{(2)} \sqrt{\pi_{i,j}^{(2)}} \right) = \sigma^2_{u(2)} \sum_{j \in unit(i)} \left(\sqrt{\pi_{i,j}^{(2)}} \right)^2 = \sigma^2_{u(2)} \\
& \text{cov}\left(\sum_{j \in unit(i_1)} u_j^{(2)} \sqrt{\pi_{i_1,j}^{(2)}}, \sum_{j \in unit(i_2)} u_j^{(2)} \sqrt{\pi_{i_2,j}^{(2)}} \right) = \sigma^2_{u(2)} \\
& \times \sum_{\substack{j \in unit(i_1) = \\ j \in unit(i_2)}} \sqrt{\pi_{i_1,j}^{(2)} \pi_{i_2,j}^{(2)}} \\
& \sum_{j \in unit(i)} \pi_{i,j}^{(2)} = 1
\end{aligned}
\tag{13.9}
$$

Unlike the multiple membership model the level 2 variation is still $\sigma^2_{u(2)}$ since we are here assuming that students actually belong to just one school, so that the true level 2 variation is unchanged. Hill and Goldstein (1998) give further details and an example. In the extreme situation where there is complete ignorance about membership we would set $\pi_{i,j}^{(2)} = 1/J$, where J is the number of level 2 units. In this case the covariance between two such students becomes $\sum_{j=1}^{J} \sigma^2_{u(2)}(1/\sqrt{J})^2 = \sigma^2_{u(2)}$ just as in the standard variance components model, and, for this variance components case, is equivalent to an arbitrary assignment of each student to a single school. Estimation for models such as (13.9) is similar to that for multiple membership models.

An alternative estimation procedure is to treat the unit membership probabilities as prior probabilities in a Bayesian analysis. This will then involve an extra MCMC step where, each level 1 unit is randomly allocated to a level 2 unit according to that level 1 unit's set of prior probabilities. The remaining MCMC steps remain the same.

We can have models which are mixtures of multiple membership structures and missing identifications. For example, in a study of examination results we may know that a student has moved school at a particular time but not know the identity of the previous school. In this case, we would specify a set of identification probabilities for the previous school which would then be multiplied by the multiple membership weight for that period of schooling.

Another application of these models is where the identification information is missing, at least partly, by design. Thus, for example, the release of survey or census data is typically subject to confidentiality safeguards and in particular this often implies the removal of locality identifications. In such cases, data could be provided where, for each actual locality, units would be assigned to it from the complete set of all localities with known probabilities. Thus, attached to each lowest level unit in a locality would be a set of probabilities of belonging to each member of the full set of localities. In practice, only a small number of such probabilities would be nonzero for each lowest level unit. If the probabilities are suitably chosen, this could satisfy confidentiality requirements since the actual locality would only be known probabilistically. The data could then be analysed using missing identification models since the probabilities are fully known for each unit.

Appendix 13.1 MCMC estimation for multiple membership models

The sampling steps for a multiple membership classification are the same as for a simple hierarchical model or a cross classification except for sampling the residuals (see Appendix 2.5).

Consider the model given in Section 13.2:

$$\begin{aligned} y_i &= (X\beta)_i + \sum_{j \in school(i)} w_{i,j}^{(2)} u_j^{(2)} + e_i \\ u_j^{(2)} &\sim N(0, \sigma^2_{u(2)}),\ e_i \sim N(0, \sigma_e^2) \\ j &= 1, \ldots J,\ i = 1, \ldots N \end{aligned} \tag{13.1.1}$$

The posterior distribution, the likelihood times the prior, is given by

$$\begin{aligned} p(u_j^{(2)}|y, \sigma_u^2, \sigma_e^2) \propto & \left(\frac{1}{\sigma_e^2}\right)^{n_j^*/2} \prod_{i=1}^{n_j^*} \exp\left[-\frac{1}{2\sigma_e^2}(y_i - ((X\beta)_i + u_{i,-j}^{(2)}) - w_{i.j}^{(2)} u_j^{(2)})^2\right] \\ & \times \left(\frac{1}{\sigma_u^2}\right)^{1/2} \exp\left[-\frac{1}{2\sigma_u^2}(u_j^{(2)})^2\right] \\ u_{i,-j}^{(2)} = & \sum_{\substack{k \in school(i) \\ k \neq j}} w_{i,k}^{(2)} u_k^{(2)} \end{aligned} \tag{13.1.2}$$

where n_j^* is the number of level 1 units that have nonzero weights for the j-th school and the product in (13.1.2) is with respect to this set of level 1 units. Thus, we need to adjust for the remaining school residuals when sampling that for school j. For a model with several random effects step 2 in Appendix 2.5 is modified to give for the means and covariance matrix of the multivariate normal posterior distribution

$$\begin{aligned} \hat{u}_j &= \left[\sum_{i=1}^{n_j^*} (w_{i.j}^{(2)})^2 Z_i^{(2)^T} Z_i^{(2)} + \sigma_e^2 \Omega_u^{-1}\right]^{-1} \left[\sum_{i=1}^{n_j^*} w_{i.j}^{(2)} Z_i^{(2)^T} (y_i - ((X\beta)_i + u_{i,-j}^{(2)}))\right] \\ \hat{D} &= \sigma_e^2 \left[\sum_{i=1}^{n_j^*} (w_{i.j}^{(2)})^2 Z_i^{(2)^T} Z_i^{(2)} + \sigma_e^2 \Omega_u^{-1}\right]^{-1} \end{aligned} \tag{13.1.3}$$

For models with several classifications we include the current random effect estimates for these classifications with the term $(X\beta)_i + u_{i,-j}^{(2)}$. Note that in the remaining steps $(X\beta)_i + u_{school(i)}^{(2)}$ in the standard model is replaced by $(X\beta)_i + \sum_{j \in school(i)} w_{i,j}^{(2)} u_j^{(2)}$.

14

Measurement errors in multilevel models

14.1 A basic measurement error model

Many measurements are made with substantial error components, especially in the social and biological sciences. Thus, if the measurement were to be repeated we would not expect always to get an identical result. In some cases, such as the measurement of individual height or weight, the errors may be so small that they can safely be ignored in practice. In other cases, for example for educational tests and attitude measures, this usually will not be true and a failure to ignore errors may lead to incorrect inferences. Fuller (2006) provides a comprehensive account of methods for dealing with measurement errors and this chapter extends some of those procedures to the multilevel model. We shall first discuss moment-based estimators and then set out a general framework based upon MCMC methods.

A basic model for measurement errors in a 2-level continuous response model for the h-th explanatory variable and for the response is as follows.

$$\begin{aligned} y^o_{ij} &= y_{ij} + q_{ij} \\ x^o_{hij} &= x_{hij} + m_{hij} \end{aligned} \tag{14.1}$$

The superscript o indicates the observed measurement. We allow measurement errors in explanatory and response variables, but will start by considering only those for the latter. We also allow the measurement errors to be correlated. We can think of the 'true' measurements as being the expected values of repeated measurements of the same unit where the measurement errors are independent and are also independent of the true values. We define the *reliability* of the h-th explanatory variable

$$R_h = \sigma^2_{hx}/\sigma^2_{hx^o} = (\sigma^2_{hx^o} - \sigma^2_{hm})/\sigma^2_{hx^o} \tag{14.2}$$

Multilevel Statistical Models: 4th Edition Harvey Goldstein

that is the variance of the true values divided by the variance of the observed values.

This immediately raises two problems. When we are measuring such things as attitudes or educational achievement, we cannot carry out repeat measurements to obtain estimates of the σ^2_{hm} because the measurement errors cannot be assumed to be independent. Another way of viewing this is to say that the process of measurement itself has changed the individual being measured, so that the underlying true value has also changed.

The second problem is that we have to define a suitable population. The definition of reliability is population dependent, so that for example, if the measurement error variance σ^2_{hm} remains constant but the population heterogeneity of the true values increases then the reliability will increase. Thus, the reliability may be lower within more homogeneous population subgroups, defined by social status say, than in the population as a whole. In particular, the reliability of a test score may be smaller within level 2 units, say schools, than across all students.

In this chapter, we shall assume that the variances and covariances of the measurement errors are known; or rather, perhaps, that suitable estimates exist, and where possible, with estimates of their precision. At the very least, sensitivity analyses can be carried out, trying different values. The topic of measurement error estimation is a complex one and there are in general no simple solutions, except where the assumption of independence of errors on repeated measuring can be made. The common procedure, especially in education, of using 'internal' measures based upon correlational patterns of test or scale items, is unsatisfactory for a number of reasons and may often result in reliability estimates that are too high. Ecob and Goldstein (1983) discuss these and propose some alternative estimation procedures. McDonald (1985) and other authors discuss the exploration and estimation of measurement error variances within a structural equation model, and these have much in common with the suggestions of Ecob and Goldstein (1983).

14.2 Moment-based estimators

All the models in this section assume that the measurements that contain measurement errors do not have random coefficients. (Details are given in Appendix 14.1.) The random coefficient case will be dealt with when we discuss MCMC estimation.

14.2.1 Measurement errors in level 1 variables

We use a two level model to show how measurement errors can be incorporated into an analysis. We write

$$y_{ij} = (X\beta)_{ij} + (z_u u)_j + (z_e e)_{ij} \tag{14.3}$$

and assume that it is this true model for which we wish to make estimates. In some situations, for example where we wish simply to make a prediction for a

response variable based upon observed values, then it is appropriate to treat these without correcting for measurement errors. If we wish to understand the nature of any underlying relationships, however, we would normally require estimates for the parameters of the true model.

For the observed variables (14.3) gives

$$y^o_{ij} = q_{ij} - (m\beta)_{ij} + (X^o\beta)_{ij} + (z_u u)_j + (z_e e)_{ij} \tag{14.4}$$

In Appendix 14.1 we show that the fixed effects can be estimated by

$$\begin{aligned} \hat{\beta} &= \hat{M}^{-1}_{xx}\hat{M}_{xy} \\ \hat{M}_{xx} &= X^{o^T}\hat{V}^{-1}X^o - C_{\Omega_1} \\ C_{\Omega_1} &= \left\{\sum_i \sigma^{ii}\sigma^i_{(h_1,h_2)m}\right\} \end{aligned} \tag{14.5}$$

where $\sigma^i_{(h_1,h_2)m}$ is the covariance between the measurement errors for explanatory variables h_1, h_2 for the i-th level 1 unit. The last expression in (14.5) is a correction matrix for the measurement errors and has elements which are weighted averages of the covariances of the measurement errors for each level over all the level 1 units in the sample with the weights being the diagonal elements of V^{-1}. In variance component models this is a simple average over the level 1 units, and in the common case where the covariance matrix of the measurement errors is assumed to be constant over level 1 units, we have

$$C_{\Omega_1} = tr(V^{-1})\Omega_{1m}, \quad \Omega_{1m} = \{\sigma_{(h_1,h_2)m}\} \tag{14.6}$$

An approximation to the covariance matrix of the estimates is given in Appendix 14.1 as is an expression for the estimation of the random parameters. For the constant measurement error covariance case with no measurement errors in the response variable, this covariance matrix is given by

$$\begin{aligned} &\hat{M}^{-1}_{xx}(X^{o^T}\hat{V}^{-1}X^o + X^{o^T}\hat{V}^{-2}T_{1m}X^o)\hat{M}^{-1}_{xx} \\ &T_{1m} = (\hat{\beta}^T\Omega_{1m}\hat{\beta})I_{(n)} \end{aligned} \tag{14.7}$$

and in the estimation of the random parameters the term T_{1m} is subtracted from $\tilde{Y}\tilde{Y}^T$ at each iteration. In some applications we may wish to allow the measurement error variances to vary as a function of explanatory variables and these will then become incorporated in C_{Ω_1}.

14.2.2 Measurement errors in higher level variables

Where variables are defined at level 2 or above and have measurement errors, we obtain analogous results (detailed in Appendix 14.1). Thus the correction term to be

used in addition to C_{Ω_1} with a constant measurement error covariance matrix in a 2-level model is

$$C_{\Omega_2} = \left(\sum_j J_{n_j}^T V_j^{-1} J_{n_j} \right) \Omega_{2m} \tag{14.8}$$

where J_n is a vector of ones of length n and V_j is the j-th block of V.

A case of particular interest is where the level 2 variable is an aggregation of a level 1 variable. Woodhouse *et al*. (1996) consider this case in detail and give detailed derivations. Consider the case where we have a level 2 variable which is the mean of a level 1 variable

$$x_{1,j}^o = \frac{1}{n_j} \sum_i x_{1ij}^o$$

The variance over the whole sample is therefore given by

$$\begin{aligned} \operatorname{var}(x_{1,j}^o) &= n_j \frac{\operatorname{var}(x_{1ij}^o)}{n_j^2} + n_j(n_j - 1) \frac{\operatorname{cov}(x_{1ij}^o x_{1i'j}^o)}{n_j^2} \\ &= \frac{1}{n_j} \sigma_{(1)}^2(x_1^o) + \frac{n_j - 1}{n_j} \sigma_{(2)}^2(x_1^o) \end{aligned} \tag{14.9}$$

where we assume constant variances and covariances within level 2 units for the x_{1ij}^o. The number of level 1 units actually measured in the j-th level 2 unit is n_j out of a total number of units N_j. Straightforward estimates of the parameters can be obtained by carrying out a variance components analysis with the variables x_{1ij}^o as responses, fitting only the overall mean in the fixed part, so that the covariance is the level 2 variance estimate.

For the true values we have an analogous result where now we consider the variance of the mean of the true values for *all* the level 1 units in each level 2 unit. There are, in effect, two sources of error in $x_{1,j}^o$. There is the error inherent in the level 1 measurement x_1^o which is averaged across the level 1 units in each level 2 unit and there is the sampling error which occurs when $n_j < N_j$, that is not all the units in the level 2 unit are measured. Thus, the true value is the average for all the level 1 units in each level 2 unit of the true level 1 measurements. Since the measurement errors are assumed independent we have

$$\operatorname{var}(x_{1.j}) = \frac{1}{N_j} \sigma_{(1)}^2(x_1) + \frac{N_j - 1}{N_j} \sigma_{(2)}^2(x_1). \tag{14.10}$$

This gives us the following expression for the required measurement error variance for the aggregated variable

$$\sigma_{1.m}^2 = \left(\frac{1}{n_j} - \frac{R_1}{N_j} \right) \sigma_{(1)}^2(x_1^o) - \left(\frac{1}{n_j} - \frac{1}{N_j} \right) \sigma_{(2)}^2(x_1^o) \tag{14.11}$$

where the reliability R_1 is estimated from the level 1 variation.

If both the level 1 observed variable and its aggregate are included as explanatory variables then clearly their measurement errors are correlated and the correlation is given by

$$\frac{1 - R_1}{n_j}\sigma^2_{(1)}(x_1).$$

In the expressions for the correction matrices, we have considered the separate contributions from levels 1 and 2. Where there is a 'cross-level' correlation between measurement errors as above then we add the level 1 variable to Ω_{2m} using (14.11) for the covariance together with a zero variance. The measurement error variance for the level 1 explanatory variable becomes a component of Ω_{1m}. A detailed derivation of these results is given by Woodhouse *et al*. (1996).

14.3 A 2-level example with measurement error at both levels

We use the Junior School Project data reading score at the age of 11 years as our response with the 8-year mathematics score as predictor, fitting also social class (Non-manual and Manual) and gender. There are 728 students in 48 schools in this analysis. The scores at ages 11 and 8 have been transformed to have a standard normal distribution, as in Table 2.7. We shall allow for measurement errors in both the test scores.

In the original analyses of these data (Mortimore *et al*., 1988), reliabilities are not given, and anyway, for the reasons given above, are unlikely to be well estimated. For the purpose of our analyses, we investigate a range of reliabilities from 0.8 to 1.0 to study the effect of introducing increasing amounts of measurement error.

It can be seen in Table 14.1 that the inferences about the fixed parameters and the level 1 variance and variance partition coefficient change markedly in moving from an assumption of zero measurement error to a reliability of 0.8. The increase in the variance partition coefficient reflects the fact that it is only the level 1 variance which decreases as the reliability decreases. The difference between the children from non-manual and manual backgrounds is reduced considerably as the reliability decreases.

We now look at the effect of adjusting additionally for measurement error in the response variable. To illustrate this we look at the effects on the individual parameters for a range of values for the reliabilities of both response and explanatory variables.

As shown in Table 14.2, as the response variable reliability decreases, so does the level 1 variance estimate. Likewise, as the reliability of the 8-year score decreases the level 1 variance decreases. The combined effect of both reliabilities being 0.8 produces a variance that is a quarter of the estimate that assumes no unreliability. When both the reliabilities reach the value of 0.7 the level 1 variance decreases to zero! By contrast, the level 2 variance is hardly altered. For the coefficient of

Table 14.1 11-year normalised mathematics score related to normalised 8-year score, gender and social class for different 8-year score level 1 reliabilities; adjusting for measurement errors in the 8-year score.

Parameter	A ($R_1 = 1.0$) Estimate (s.e.)	B ($R_1 = 0.9$) Estimate (s.e.)	C ($R_1 = 0.8$) Estimate (s.e.)
Fixed			
Intercept	0.13	0.10	0.07
8-year score	0.67 (0.026)	0.75 (0.030)	0.86 (0.035)
Gender	−0.048 (0.050)	−0.048 (0.050)	−0.047 (0.052)
Non-manual	0.14 (0.06)	0.10 (0.06)	0.06 (0.06)
Random			
σ_u^2	0.080 (0.023)	0.081 (0.024)	0.081 (0.024)
σ_e^2	0.422 (0.023)	0.373 (0.023)	0.310 (0.025)
Variance partition coefficient	0.16	0.18	0.21

the 8-year score and social class the greatest change is with the reliability of the 8-year score.

As the reliability decreases so the strength of the relationship with 8-year score increases, while the social class difference decreases substantially. The gender difference is changed very little.

Clearly, the requirement of a positive level 1 variance implies particular lower bounds on the reliabilities and measurement error variances, and underlines the importance of obtaining good estimates of these parameters or at least a range of reasonable estimates. We also have restrictions on the correlations between measurement errors. Consider the case of two explanatory variables with measurement error, and suppose for simplicity that they have the same observed variance equal to 1 and the same reliability, R. Let us also suppose that their measurement errors have a correlation of ρ_m and that the correlation between the observed variables is ρ_o. Now, we require that the correlation between the true values lies between –1 and 1 and this implies

$$\frac{\rho_o + R}{1 - R} > \rho_m > \frac{\rho_o - R}{1 - R}.$$

Thus, say, if $R = 0.7$ and $\rho_o = 0.8$ then we require $\rho_m > 0.33$.

In our example, the range of variance partition coefficient values, from 16 % to 21 %, also indicates that we need to take care in interpreting small values of such coefficients without adjusting for measurement error. This has important implications for 'school effectiveness' studies (Goldstein, 1997) where ignoring measurement errors can be expected to lead to underestimation of school effects.

Table 14.2 Parameter estimates for values of explanatory and response variables.

8-year score		Response reliability		
		1.0	0.9	0.8
8-year score reliability	1.0	0.67	0.67	0.67
	0.9	0.75	0.75	0.75
	0.8	0.86	0.86	0.87
Gender		Response reliability		
		1.0	0.9	0.8
8-year score reliability	1.0	−0.048	−0.048	−0.047
	0.9	−0.048	−0.048	−0.047
	0.8	−0.047	−0.046	−0.046
Non-manual		Response reliability		
		1.0	0.9	0.8
8-year score reliability	1.0	0.14	0.14	0.15
	0.9	0.10	0.10	0.11
	0.8	0.06	0.06	0.06
Level 2 variance		Response reliability		
		1.0	0.9	0.8
8-year score reliability	1.0	0.080	0.079	0.078
	0.9	0.081	0.080	0.079
	0.8	0.081	0.080	0.079
Level 1 variance		Response reliability		
		1.0	0.9	0.8
8-year score reliability	1.0	0.422	0.324	0.225
	0.9	0.373	0.274	0.176
	0.8	0.310	0.211	0.112

14.4 Multivariate responses

To model multivariate data, we specify a dummy (0,1) variable for each response and corresponding interactions with other explanatory variables. Then $C_{\Omega_1}, C_{\Omega_2}$ in (14.5) and (14.8) are modified so that for each level 1 or level 2 unit, the covariance between measurement errors is set to zero when either of the corresponding dummy variables is zero and likewise for the variances. This is equivalent to specifying the same covariance matrix of measurement errors for each set of explanatory variables corresponding to a response variable, with no covariances across these sets. For the response variables we likewise specify the separate measurement error variances for each one using the general procedures in Appendix 14.1.

14.5 Nonlinear models

Consider the 2-level model (9.3) in Chapter 9 where there are measurement errors in the explanatory variables for the fixed part of the model. In this case we can obtain an approximate analysis by using the observed values in the updating formulae and replacing the measurement error covariances in (14.5) by

$$(f'_{(i)})^2 \sigma^i_{(h_1,h_2)m} \tag{14.12}$$

where $f'_{(i)}$ is the first differential of the nonlinear function for the i-th level 1 unit with a corresponding expression for level 2 measurement errors. The derivation of (14.12) is given in Appendix 14.1.

14.6 Measurement errors for discrete explanatory variables

Assume that we have a categorical explanatory variable with r categories. We shall consider only a single such variable, since multiple variables can in principle be handled by considering the p-way table based upon them as a single vector. In practice, it will often be reasonable to assume that their measurement errors are uncorrelated so that they can be considered separately. Likewise, for now, we shall assume that measurement errors in discrete explanatory variables are uncorrelated with those in continuous variables. The following derivations parallel those given by Fuller (2006, Section 3.4). We consider only level 1 explanatory variables, but the extension to higher levels follows straightforwardly. We shall consider more general models for discrete variables when we look at MCMC estimation.

Let $A_{(i)}$ be a row vector for the *i-th* level 1 unit containing a one for the category which is observed and zeros elsewhere. Let k_{mn} be the probability that a level 1 unit with true category n is observed in category m. We write

$$K = \{k_{mn}\}, \quad \text{where } K_m \text{ is the } m - th \text{ column of } K$$

and define

$$x^{o^T}_{(i)} = K^{-1} A^T_{(i)} \tag{14.13}$$

If $x_{(i)}$ is the true value we write

$$A_{(i)} = x_{(i)} + \varepsilon_{(i)}, \quad E(A_{(i)}|x_{(i)}) = x_{(i)} K^T$$

We also write

$$x^o_{(i)} - x_{(i)} = m_{(i)}$$

so that

$$E(m_{(i)}|x_{(i)}) = 0$$

which gives the familiar form for the errors in variables model where the unknown true value $x_{(i)}$ is uncorrelated with the measurement error. The $x^o_{(i)}$ become the new set of observed values and interest is in the regression on the true category values $x_{(i)}$. The vector $x_{(i)}$ consists of a single value of one and the remainder zero. We have

$$\text{cov}(A^T_{(i)}|x_{(i)} = l_m) = \Sigma_{(i)(m)} = diag(K_m) - K_m K_m^T$$

where l_m is an r-dimensional vector with 1 in the m-th position and zeros elsewhere. For the i-th level 1 unit define

$$\Omega_{(i)m} = \text{cov}(m^T_{(i)}|x_{(i)} = l_m) = K^{-1}\Sigma_{(i)(m)}K^{-1^T} \tag{14.14}$$

and we use as our estimate of the covariance matrix of measurement errors the matrix in (14.14) conditional on the observed $A_{(i)}$.

$$\hat{\Omega}_{(i)m} = \Omega_{(i)m}\left[\frac{P(x_{(i)} = l_m)}{P(A_{(i)} = l_m)}\right] \tag{14.15}$$

The term in square brackets can be estimated as follows. If μ_A,μ_x are the observed and true vectors of probabilities for the categories, then

$$\mu_x = K^{-1}\mu_A$$

and given the sample estimate of μ_A we can estimate μ_x. The estimate given by (14.15) is then used as in the case of continuous explanatory variables measured with error. In the general model the number of explanatory variables will generally be one less than the number of categories, with one of the categories chosen as the base and omitted.

In practice, the matrix of probabilities K, is normally assumed constant but can itself depend on further explanatory variables. Often we will not have a good estimate of it, and we may need to make some simplifying assumptions. In the case of a binary variable, it may be possible to assume equal misclassification probabilities, in which case only a single value needs to be determined, and in practice a range of values can be explored. (We discuss this further in Appendix 14.1.)

14.7 MCMC estimation for measurement error models

Appendix 14.1 sets out the steps involved in modelling measurement error data using MCMC. It involves adding extra steps to the existing Gibbs or MH algorithms

Table 14.3 11-year normalised mathematics score related to normalised 8-year score, gender and social class for different 8-year score level 1 reliabilities; adjusting for measurement errors in the 8-year score. MCMC estimates.

Parameter	A ($R_1 = 1.0$) Estimate (s.e.)	B ($R_1 = 0.9$) Estimate (s.e.)	C ($R_1 = 0.8$) Estimate (s.e.)
Fixed			
Intercept	0.14	0.13	0.13
8-year score	0.67 (0.027)	0.76 (0.04)	0.90 (0.04)
Gender	−0.039 (0.050)	−0.039 (0.050)	−0.041 (0.051)
Non-manual	0.15 (0.06)	0.15 (0.06)	0.15 (0.06)
Random			
σ^2_{u0}	0.093 (0.023)	0.091 (0.026)	0.091 (0.026)
σ_{u01}	−0.020 (0.010)	−0.018 (0.010)	−0.016 (0.011)
σ^2_{u1}	0.014 (0.008)	0.012 (0.007)	0.013 (0.007)
σ^2_e	0.413 (0.023)	0.362 (0.023)	0.270 (0.026)

and assuming prior distributions for the true scores and measurement errors. The REALCOM software package (REALCOM, 2008) will fit these models and some can also be fitted in MLwiN version 2.10 onwards.

The principal extension to the standard MCMC algorithm is to insert an extra step that involves sampling the true values given the observed values and the current parameter estimates. The algorithm can allow for correlated measurement errors that can vary across level 1 units and will also handle measurement errors in the response. Its principal advantage is that it will allow measurement errors in explanatory variables for any random effects and it will also allow informative priors for the measurement errors that can reflect uncertainty in their estimated values. Both continuous and discrete variables can be treated. Goldstein *et al*. (2008) give details with examples.

We now apply this to the data in Table 14.1, including also a random coefficient.

As Table 14.3 shows, the results are similar to those obtained for the variance components model when increasing amounts of measurement error are introduced, with a somewhat greater effect on the level 1 variance estimate and now no effect on the social class coefficient.

As outlined in Appendix 14.1, we can also extend the MCMC algorithm to include correlations among measurement errors, to include misclassifications for discrete variables, and also a latent normal model for categorical variables.

Appendix 14.1 Measurement error estimation

14.1.1 Moment based estimators for a basic 2-level model

Consider a 2-level model and write for the response and explanatory variables

$$\begin{aligned} y^o_{ij} &= y_{ij} + q_{ij} \\ x^o_{hij} &= x_{hij} + m_{hij} \\ \text{cov}(q_{ij}q_{i'j}) &= \text{cov}(m_{hij}m_{hi'j}) = 0 \\ E(q_{ij}) &= E(m_{hij}) = 0 \\ \text{cov}(m_{h_1ij}m_{h_2ij}) &= \sigma^i_{(h_1h_2)jm} \end{aligned} \tag{14.1.1}$$

for the h-th explanatory variable with measurement error vector m_h and with q as the measurement error vector for the response. We use the superscript o for the observed values. Each level 1 unit may have its own set of measurement error variances. Where we have a level 2 explanatory variable, then the measurement error is constant within a level 2 unit.

We write the 'true' model in the general form

$$y_{ij} = (x\beta)_{ij} + (z_u u)_j + (z_e e)_{ij} \tag{14.1.2}$$

which gives the model for the observed variables as

$$\begin{aligned} y^o_{ij} &= q_{ij} - (m\beta)_{ij} + (x^o\beta)_{ij} + (z_u u)_j + (z_e e)_{ij} \\ m &= \{m_h\} \end{aligned} \tag{14.1.3}$$

For the true values write

$$\begin{aligned} M_{xx} = x^T V^{-1} x, \quad \text{M}_{\text{xy}} = x^T V^{-1} y \\ \hat{\beta} = M^{-1}_{xx} M_{yy} \end{aligned} \tag{14.1.4}$$

Now

$$\begin{aligned} {x^o}^T V^{-1} x^o &= (x+m)^T V^{-1}(x+m) \\ &= x^T V^{-1} x + m^T V^{-1} x + x^T V^{-1} m + m^T V^{-1} m \end{aligned} \tag{14.1.5}$$

so that

$$E({x^o}^T V^{-1} x^o) = x^T V^{-1} x + E(m^T V^{-1} m) \tag{14.1.6}$$

If we further assume that q and m are uncorrelated then we have

$$E({x^o}^T V^{-1} y) = x^T V^{-1} y \tag{14.1.7}$$

Thus, to estimate the fixed parameters we require $E(m^T V^{-1} m)$ and we now consider how to obtain this for measurement errors at both level 1 and level 2 using essentially moment based estimators.

14.1.2 Parameter estimation

For errors of measurement in level 1 units the (h_1, h_2) element of $E(m^T V^{-1} m)$ is

$$\begin{aligned} &\sum_{i=1}^{N} \sigma^{ii} \sigma^{i}_{(h_1 h_2) jm} \\ &\text{with} \quad C_{\Omega_1} = \left\{ \sum_i \sigma^{ii} \sigma^{i}_{(h_1 h_2) jm} \right\} \end{aligned} \tag{14.1.8}$$

where N is the total number of level 1 units. In the case where each level 1 unit has the same covariance matrix of measurement errors we have

$$C_{\Omega_1} = tr(V^{-1}) \Omega_{1m}, \quad \Omega_{1m} = \{\sigma_{(h_1 h_2) m}\} \tag{14.1.9}$$

For errors of measurement in level 2 explanatory variables, we have

$$C_{\Omega_2} = \sum_j (J_{(1,n_j)} V_j^{-1} J_{(n_j,1)}) \Omega_{2jm} \tag{14.1.10}$$

where Ω_{2jm} is the covariance matrix of measurement errors for the j-th level 2 block, and $J_{(r,s)}$ is a (r x s) matrix of ones. In Chapter 14 we discuss how to obtain the $\Omega_{2jm}$for level 2 variables which are aggregates of level 1 variables.

For the measurement error corrected estimate of the fixed coefficients, we have

$$\hat{M}_{xx} = M_{XX} - C_{\Omega_1} - C_{\Omega_2} \tag{14.1.11}$$

For the random component based upon the model with observed variables, write the residual $v_{ij} = (z_u u)_j + (z_e e)_{ij} + q_{ij} - (m\beta)_{ij}$, $v = \{v_{ij}\}$which gives

$$\begin{aligned} &E(vv^T) = V + \bigoplus_{ij} \sigma^2_{ijq} + T_1 + T_2 \\ &T_1 = \bigoplus_{ij} (\hat{\beta}^T \Omega_{1ijm} \hat{\beta}), \quad T_2 = \bigoplus_{j} (\hat{\beta}^T \Omega_{2jm} \hat{\beta}) J_{(n_j, n_j)} \end{aligned} \tag{14.1.12}$$

where σ^2_{ijq} is the measurement error variance for the ij-th response measurement. Thus the quantity $\bigoplus_{ij} \sigma^2_{ijq} + T_1 + T_2$ should be subtracted from the sum of products matrix $\tilde{Y}\tilde{Y}^T$ at each iteration, when estimating the random parameters.

The covariance matrix of the estimated fixed coefficients is given by

$$\begin{aligned} &\hat{M}_{xx}^{-1}(x^{o^T}V^{-1}x^o + x^{o^T}V^{-1}QV^{-1}x^o + x^{o^T}V^{-2}[T_1 + T_2]x^o)\hat{M}_{xx}^{-1} \\ &Q = \bigoplus_{ij} \sigma_{ijq}^2 \end{aligned} \tag{14.1.13}$$

This expression ignores any variation in the estimation of the measurement error variance itself, although Goldstein (1986) includes terms for this.

14.1.3 Random coefficient for explanatory variables measured with error

The above expressions assume that the coefficients of variables with measurement error are not random. Where such coefficients are random the above formulae do not apply, and in particular $m^T V^{-1} m$ has measurement errors in all its components. Woodhouse (1998) discusses this problem in detail and suggests that moment-based estimators are not really feasible. This problem does not arise when we use MCMC estimation as described below.

14.1.4 Nonlinear models

Consider first the case where just the fixed part explanatory variables have measurement errors at level 1 in the single component 2-level nonlinear model for the i-th level 1 unit

$$y_{(i)} = f_{(i)}(x^o\beta + random)$$

which yields the linearisation

$$y_{(i)} - \{f_{(i)}(x^o\beta) - \sum_k \beta_{k,t}x^*_{(i)k}\} = \sum_k \beta_{k,t+1}x^*_{(i)k} + random \tag{14.1.14}$$

where the explanatory variables are the observed measurements and the coefficients are the required ones corrected for measurement error and $x^*_{(i)k} = f'_{(i)}x^o_{(i)k}$. Consider the expansion of $f_{(i)}$ for the measurement error terms, to a first order approximation,

$$f_{(i)} = f_{(i),m_k=0} + \sum_k f'_{(i),u_k=0}m_{(i)k}\beta_{k,t} \tag{14.1.15}$$

Thus, we can use the observed explanatory variables with measurement error as an approximation to the use of the true values in the updating formulae, with $(f'_{(i)})^2\sigma^i_{(h_1,h_2)m}$ replacing $\sigma^i_{(h_1,h_2)m}$ in (14.5).

14.1.5 MCMC estimation for measurement error models: continuous variables

Suppose we have p explanatory variables containing measurement error and q that do not. We write the model as:

$$\begin{aligned} y_{ij} &= [X_{1ij}(\beta_1 + Z_{1ij} \cdot U_{1j})] + [X_{2ij}(\beta_2 + Z_{2ij} \cdot U_{2j})] + e_{ij} \\ \beta^T &= \{\beta_1^T, \beta_2^T\}, Z^T = \{Z_1^T, Z_2^T\}, U = \{U_1, U_2\} \end{aligned} \tag{14.1.16}$$

where the explanatory variable matrix of true values for those with measurement error is X_1(N x p) and that for those without error is X_2(N x q). For the random part, explanatory variables Z_1, Z_2 are indicator vectors of dimensions (p x 1) and (q x 1), with ones or zeros, so that the dot (Hadamard) product with the level 2 residuals selects the explanatory variables for the random part of the model – assuming that these are a subset of the fixed part explanatory variables. Using the notation of Goldstein *et al*. (2008) we have

$$X_1^O \sim MVN(X_1, \Omega_m), \quad X_1 \sim MVN(\theta, \Omega_\phi) \tag{14.1.17}$$

where X_1^O is the matrix of observed values and Ω_m is the covariance matrix of measurement errors, initially assumed to be common to all level 1 units, θ is the mean vector and Ω_ϕ is the covariance matrix of the true values of X_1. MCMC estimation is used to obtain the following posterior distributions.

$$\begin{aligned} p(\theta|X_1, \Omega_\phi) &\sim MVN(\hat{\theta}, \hat{V}_\theta), \quad \hat{\theta} = \bar{X}_1^0, \hat{V}_\theta = \Omega_\phi/N \\ p(\Omega_\phi^{-1}|X_1, \theta) &\sim Wishart(N - 3, [(X_1 - \hat{\theta})^T(X_1 - \hat{\theta})]^{-1}) \end{aligned} \tag{14.1.18}$$

where N is the number of level 1 units. Since θ is a row vector of means we assume a uniform prior for θ. We can also choose, and then sample from, a prior distribution for the measurement error covariance matrix. An obvious choice is $p(\Omega_m^{-1}) \sim Wishart(\delta_p, \delta_p S_m)$ and we might wish to assume a minimally informative choice where the degrees of freedom δ_p is equal to the order of the matrix and S_m is a covariance matrix chosen on the basis of existing evidence or on theoretical grounds.

In practice, there will often be so much uncertainty about $\Omega_m$that it may be illuminating to select a range of values for S_m and examine the effects conditional on these choices, in the spirit of sensitivity analysis. For each choice we may also choose a prior distribution for Ω_m.

For Ω_ϕ we could also assume a general inverse Wishart prior, but it is not clear what parameters we should use, so we have assumed a uniform prior here by setting the 'degrees of freedom' parameter of the Wishart distribution in (14.1.18) to N-3.

The sampling for the fixed parameters, β, the residuals, measurement error covariance matrix (conditional on measurement error estimates), level 2 covariance matrix and level 1 variance, conditional on the X_1, X_2 and given priors, is as in the standard case.

For sampling the X_1 we write

$$p(X_1|y, X_1^O; \beta, U, \sigma_e^2, \Omega_\phi, \Omega_m) = p(y|X_1; \beta, U, \sigma_e^2) \times p(X_1^O|X_1, \Omega_m)p(X_1|\Omega_\phi) \tag{14.1.19}$$

which leads to the following sampling for each row of X_1.

$$\begin{aligned} X_{1ij} &\sim MVN(\hat{X}_{1ij}, \hat{V}_{ij}) \quad \text{where} \\ \hat{V}_{ij} &= \left[\frac{(\beta_1 + Z_1 \cdot U_{1j})(\beta_1 + Z_1 \cdot U_{1j})^T}{\sigma_e^2} + \Omega_m^{-1} + \Omega_\phi^{-1}\right]^{-1} \\ \hat{X}_{1ij} &= \hat{V}_{ij}\left[\frac{(\beta_1 + Z_1 \cdot U_{1j})(y_{ij} - X_{2ij}(\beta_2 + Z_2 \cdot U_{2j}))}{\sigma_e^2}\right. \\ &\quad \left. + X_{1ij}^O\Omega_m^{-1} + \theta\,\Omega_\phi^{-1}\right] \end{aligned} \tag{14.1.20}$$

where $Z \cdot U$denotes the Hadamard vector product. The level 1 residuals are obtained by subtraction. Note that in the so-called 'functional' model Ω_ϕ is zero, and this term is omitted from the expressions in (14.1.20).

In some applications the measurement error covariance matrix may vary across level 1 (or level 2) units, for example as a known function of predictor variables. In this case we simply replace Ω_m^{-1} by Ω_{mij}^{-1} in (14.1.20).

If we have measurement error in the response

$$y^O = y + e_y, \quad e_y \sim N(0, \sigma_{e_y}^2) \tag{14.1.21}$$

where, in order to ensure identification, we must have known variance $\sigma_{e_y}^2$. We apply this to the residuals using the adjusted value $\sigma_{e_y}^{*^2} = \sigma_e^2\sigma_{e_y}^2/\sigma_y^2$ and we insert the extra step to sample y_{ij} from

$$N\left[\left(\sigma_e^2 - \sigma_{e_y}^{*^2}\right)\sigma_e^{-2}\tilde{y}_{ij} + \hat{y}_{ij}, \left(\sigma_e^2 - \sigma_{e_y}^{*^2}\right)\sigma_e^{-2}\sigma_{e_y}^{*^2}\right] \tag{14.1.22}$$

where $\hat{y}_{ij}$is the predicted value and $\tilde{y}_{ij} = y_{ij}^O - \hat{y}_{ij}$.

We require that the covariance matrix of the true explanatory variables is positive definite so that having sampled the X_1, if this is not the case, then we retain the existing values.

14.1.6 MCMC estimation for measurement error models: discrete variables

We set out details or the binary case and then discuss multiple categories. We assume that the errors (misclassifications) are independent across variables.

Write the probability of observing a zero given that the true value is zero as $P_{obs}(0|0)$ and the probability of observing a one given that the true value is a zero as $P_{obs}(1|0)$, etc. Then the probability of observing a zero is $P_{obs}(0) = P_{true}(0)P_{obs}(0|0) + P_{true}(1)(P_{obs}(0|1))$ and the probability of observing a one

is $P_{obs}(1) = P_{true}(1)(1 - P_{obs}(0|1)) + P_{true}(0)(1 - P_{obs}(0|0))$ where $P_{true}(0)$, $P_{true}(1)$ are the true probabilities of a zero and one.

This gives the following values for the true (prior) probabilities

$$P_{true}(0) = \frac{P_{obs}(1|1) - P_{obs}(1)}{P_{obs}(1|1) + P_{obs}(0|0) - 1}, \quad P_{true}(1) = 1 - P_{true}(0) \tag{14.1.23}$$

Consider a normal response model. The probability for an observation that has true value zero where we *observe* a zero for the binary variable x_1 with coefficient β_1 which is assumed to have a uniform prior, is proportional to

$$L_{00} = \exp\left(-\frac{(\tilde{y})^2}{2\sigma_e^2}\right) P_{obs}(0|0)$$

and for an observed zero where the true value is one we have the probability proportional to

$$L_{01} = \exp\left(-\frac{(\tilde{y} - \beta_1)^2}{2\sigma_e^2}\right) P_{obs}(1|0)$$

where $\tilde{y}$ is the observed response minus predicted value of the response given the remaining parameters.

When a zero is observed, combining these probabilities with the priors, we select a new true value to be zero with probability

$$\frac{L_{00} P_{true}(0)}{L_{00} P_{true}(0) + L_{01} P_{true}(1)}.$$

We have corresponding results when a one is observed, namely

$$\begin{aligned} L_{10} &= \exp\left(-\frac{(\tilde{y})^2}{2\sigma_e^2}\right) P_{obs}(1|0) \\ L_{11} &= \exp\left(-\frac{(\tilde{y} - \beta_1)^2}{2\sigma_e^2}\right) P_{obs}(1|1) \end{aligned} \tag{14.1.24}$$

and we select a new true value of one with probability

$$\frac{L_{11} P_{true}(1)}{L_{11} P_{true}(1) + L_{10} P_{true}(0)}.$$

Having sampled a new set of true values we then apply the standard steps in the MCMC algorithm for the remaining parameters. For generalised linear models the only change is in the expressions for the likelihoods and if we use, for example, a probit link with binary data then there is no change except for the extra step generating a normally distributed response from the binary response.

For variables with multiple categories we can extend the above using estimates for all possible misclassification probabilities. Suppose we have a variable x with p categories. In principle, each category can be misclassified into one of the remaining *p-1* categories. This implies values for $(p-1)^2$ parameters. Even if some of these can be set to zero, for example, in the case of ordered classifications, this in general sets practical limits to the data structures that can be fitted and is likely to lead to computational problems. Instead, therefore, we propose the following scheme.

For each category we use, as in the binary case, just the misclassification probabilities $P^{(k)}_{obs}(1|0)$, $P^{(k)}_{obs}(1|1)$ for each category, where '1' indicate that we have category k and '0' that we have any other category. From this we can derive the true prior probabilities, as before. We have the true probability for category k

$$P^{(k)}_{true}(1) = \frac{P^{(k)}_{obs}(0|0) + P^{(k)}_{obs}(1) - 1}{P^{(k)}_{obs}(1|1) + P^{(k)}_{obs}(0|0) - 1}$$

We need to ensure that these lie in (0,1) and also that they sum to 1, that is

$$\sum_k \frac{P^{(k)}_{obs}(0|0) + P^{(k)}_{obs}(1) - 1}{P^{(k)}_{obs}(1|1) + P^{(k)}_{obs}(0|0) - 1} = 1$$

This latter condition is always satisfied when there are just two categories.

In the expression for the posterior acceptance probabilities, if we actually observe a value in category k we have, corresponding to (14.1.24)

$$L^{(k)}_{11} = \exp\left(\frac{(\tilde{y} - \beta_k)^2}{2\sigma_e^2}\right) P^{(k)}_{obs}(1|1), \quad L^{(k)}_{10} = \exp\left(\frac{(\tilde{y})^2}{2\sigma_e^2}\right) P^{(k)}_{obs}(1|0) \qquad (14.1.25)$$

where the superscript k refers to the k-th category and β_k is the coefficient of the dummy variable for this category – note that for the reference category this will be 0.

Thus we accept a new true value for category k with probability

$$\frac{L^{(k)}_{11} P^{(k)}_{true}(1)}{L^{(k)}_{11} P^{(k)}_{true}(1) + L^{(k)}_{10} P^{(k)}_{true}(0)}$$

If the value is not accepted then it has to be allocated to one of the remaining categories. In the binary case this is unproblematic and for the multicategory case we assign with probability proportional to the true probabilities as calculated above.

In the case where the categories are ordered, we may be able to specify the observed probabilities in terms of true probabilities for the nearest categories only, in which case it may be practical to work with all of these deriving corresponding expressions to those above.

14.1.7 A latent normal model for discrete and continuous variables with measurement errors

We shall illustrate this using the case of binary variables, but the extension to general categorical variables as set out in Chapter 7 is readily implemented.

Suppose z is our true binary (0,1) variable and y the corresponding latent normal variable $y \sim N(0, 1)$. Our corresponding observed variables are z^*, y^*. We now write $y^* = y + m$ as our measurement error basic model. We have shown how to sample $y|y^*, \ldots$. and we obtain y^* from z^* using a standard Gibbs step based upon the model

$$P(z = 1) = \int_{\alpha}^{\infty} \phi(y)dy, \quad P(z^* = 1) = \int_{\alpha}^{\infty} \phi(y^*)dy^* \tag{14.1.26}$$

Thus, we first sample y^* given z^* and the current parameter values, and then sample $y|y^*$ as in the case of continuous variables. To do this we require an estimate of σ_m^2 and if there are several variables the joint distribution of measurement error variances.

Assume without loss of generality that the observed latent variable has a standard normal distribution and we know $P_{obs}(0|0)$, $P_{obs}(0|1)$, and hence $P_{true}(1)$ and $P_{true}(0)$. The joint distribution of observed and true latent normal variables is

$$\begin{pmatrix} y^* \\ y \end{pmatrix} \sim N(0, \Omega), \quad \Omega = \begin{pmatrix} 1 & \\ 1 - \sigma_m^2 & 1 - \sigma_m^2 \end{pmatrix} \tag{14.1.27}$$

For this bivariate normal distribution, ϕ_m say, the orthant probabilities given by

$$P(obs = 1, true = 1) = \int_{0}^{\infty} \phi_m(y^*; y > 0)dy^*,$$
$$P(obs = 1, true = 0) = \int_{0}^{\infty} \phi_m(y^*; y < 0)dy^*$$

are simply functions of σ_m^2. This then allows us to estimate σ_m^2 using a suitable algorithm (see, for example, Craig, 2007) for computing these probabilities given values of σ_m^2. This becomes more difficult if we have a joint distribution of measurement errors. One possibility is to form a combined multicategory variable by combining all the categorical variables with error, using the above procedure for selecting a true (joint) category and then sampling the latent normal variables as in Chapter 7. This still assumes that the categorical and continuous errors are uncorrelated, but we can allow the continuous measurement error variances to depend on the true categories (Goldstein *et al*., 2008).

To allow fully for correlation across categorical and continuous variables, we would first need to sample y^* given z^* and the continuous observed variables with measurement errors, and then sample from the joint distribution of true values. Having sampled y, we then choose $z(=1)$ according to whether $y > \alpha$.

15

Smoothing models for multilevel data

15.1 Introduction

For many kinds of data we may wish to fit a smooth, nonparametric, function of a continuous predictor (x). For example, fitting growth curves using polynomials may perform poorly at end points where data are sparse and polynomial coefficients are determined by those parts of the continuum where data are concentrated and so may be poorly conditioned at the ends. The advantage of smooth functions (smoothers) in such situations is that they are locally influenced, as explained in Section15.3.

There is a large literature on ways of fitting such functions which for the single level case have the form

$$y_i = f(x_i) + e_i, e_i \sim N(0, \sigma_e^2) \tag{15.1}$$

and these are sometimes referred to as generalised additive models (for general introductions, see Hastie and Tibshirani, 1990; Green and Silverman, 1994). The model (15.1) can be extended by adding covariates and also random effects to handle multilevel data as will be shown.

The next section reviews the main types of smoothing estimators for single level models and this followed by various multilevel extensions with an example.

15.2 Smoothing estimators

15.2.1 Regression splines

We first look at ways of fitting Model (15.1). One idea is that a function, typically a low order polynomial, is fitted within each of a set of intervals along the line defined

Multilevel Statistical Models: 4th Edition Harvey Goldstein

by t, with joins at 'knots' and the function constrained to be smooth at these joins in terms of the equality of their derivatives. These are known as regression splines, with cubic splines the most common, having first and second derivatives that are zero. To represent these splines we require a 'basis' function on which they are defined. B splines, a standard choice (Silverman, 1986), have computational advantages but do not have a natural generalisation to handle random effects. Instead, we can use the truncated power series (also known as grafted polynomials or + functions) basis and we write a 2-level model as

$$\begin{aligned} y_{ij} &= \sum_{l=0}^{3} \beta_l x_{ij}^l + \sum_{k=1}^{q} \theta_k (x_{ij} - \xi_k)_+^3 + u_j + e_{ij} \\ (t)_+ &= \begin{cases} t, & t \geq 0 \\ 0, & t < 0 \end{cases} \end{aligned} \tag{15.2}$$

where the β_l, θ_k are regression coefficients and q is the number of knots or join points and it is automatically smooth at the knots. The order of the polynomials can be changed and we can have different order polynomials at each knot, as used by Blatchford *et al*. (2002) for modelling class size and by Pan and Goldstein (1998) for growth curve models. The main problem is deciding on the number and spacing of the knots. When there are too few knots, the estimates can be very sensitive to their placing, and evaluating all possible knot positions can be very time consuming (see however Friedman and Silverman, 1989). Other approaches related to this such as that using 'penalty functions' are discussed by Verbyla *et al*. (1999).

15.3 Smoothing splines

The other major approach is to estimate a smooth function which is defined at each data point X_i; known as smoothing splines. Consider first a simple running mean 'bin smoother' where we have a moving box width h placed symmetrically about each data point and at each of a chosen set of values $\{x\}$ we estimate the function as the mean of the y values in the box. This is equivalent to the following:

$$\begin{aligned} f(x) &= A/B \\ A &= \tfrac{1}{h} \sum_{i=1}^{n} y_i w\left(\tfrac{2(x - X_i)}{h}\right) \\ B &= \tfrac{1}{h} \sum_{i=1}^{n} w\left(\tfrac{2(x - X_i)}{h}\right) \\ w(t) &= 1 \text{ if } |t| < 1, \text{ otherwise } 0 \end{aligned} \tag{15.3}$$

For computational purposes we may first wish to centre the y values. This function, however, generally is not very smooth. Instead of a fixed width it is generally better to choose the p observations nearest to x (in terms of absolute value or symmetrically

placed) and this has the advantage of allowing the bandwidth to be a function of the local point density. Also, instead of computing the mean in each interval, we can fit a least squares (LS) line to the data in each interval and use the predicted value at *x*. This is the method used by Healy *et al*. (1988). A development of this is to use a weighted LS fit where the weights decrease with distance from *x*. It is in this sense that generally we consider a smoother to be 'locally influenced'. The weight function is typically referred to as a 'kernel'. The LOWESS smoother (Cleveland and Devlin, 1988) is a particularly useful implementation of this and procedures for choosing the span, *p/n,* are available.

General kernel smoothers use a weight function that is defined, at each point, for the whole range of the data. Thus we have

$$w(x) = \frac{c}{h} g\left(\frac{|x - X_i|}{h}\right) \tag{15.4}$$

such that, for each point, the weights sum to 1 over the whole set of points.

Note that the box width, h, or bandwidth, could vary over the range, although commonly it is assumed constant. As we shall illustrate in our example, the choice of kernel function g, which decreases as $|x - X_i|$ increases, may be important. A common choice is based on the standard normal distribution function, that is, $g(u) = \exp(-u^2/2)$. We shall also look at a quartic kernel of the form

$$g(u) = \begin{cases} (1 - u^2)^2, & u \leq 1 \\ 0 & u > 1 \end{cases}$$

The choice of bandwidth is important. If it is too narrow the function will follow the data too closely and not be very smooth, and if too wide it will be smooth but may obscure important local variation.

For all these methods we can write the predicted smoothed curve as $\hat{y} = Sy$ where S is the $n \times n$ smoother matrix whose i-th row is the set of normalised weights for the i-th point. Thus, for example, for a bin smoother this will contain values of zero for points outside the bin. The covariance matrix of the predicted curve is given by

$$\text{cov}(\hat{y}) = SS^T\sigma_e^2 \tag{15.5}$$

15.4 Semiparametric smoothing models

In general, we will want to incorporate other predictors or covariates into a model such as (15.1). We may also wish to model more than one smooth function, a case which will also occur if we have an interaction between a smooth function and a predictor variable. In the case of two smooth additive functions we can write, omitting subscripts

$$y = f_1(x_1) + f_2(x_2) + e$$

A simple 'back fitting' algorithm alternately fits both components, so that starting with, say, an initial estimate $f_1^{(1)}(x_1)$ we then estimate the second function using the adjusted response $y - f_1^{(1)}(x_1)$ and then re-estimate the first function using $y - f_2^{(1)}(x_2)$, and so on until convergence. Note that it is the function values and not the smoother matrices that are updated. The back-fitting algorithm solves the set of equations

$$\begin{pmatrix} I & S_1 \\ S_2 & I \end{pmatrix} \begin{pmatrix} f_2 \\ f_2 \end{pmatrix} = \begin{pmatrix} S_1 y \\ S_2 y \end{pmatrix}$$

with straightforward extension to the case of more than two functions; this solution can be expressed using standard results for inverting partitioned matrices. We obtain

$$\hat{y} = \{I - (I - S_2)(I - S_1 S_2)^{-1}(I - S_1)\}y = S_{12}y \tag{15.6}$$

where we require the matrix norm $||S_1 S_2|| < 1$.

We now consider a general model with a parametric component and a nonparametric component which we can write as

$$y = (X\beta) + f(t) + e \tag{15.7}$$

A penalised least squares approach leads to the following procedure, which is associated with a minimum bias.

We first obtain the smoothing matrix S. We then form

$$\begin{aligned} \tilde{X} &= (I - S)X = X - SX \\ \tilde{y} &= (I - S)y = y - Sy \\ \hat{\beta} &= (\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T \tilde{y} \\ \hat{f} &= S(y - X\hat{\beta}) \end{aligned} \tag{15.8}$$

with large sample covariance estimates

$$\text{cov}(\hat{\beta}) = \hat{\sigma}_e^2 (\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T (I - S)(I - S)^T \tilde{X} (\tilde{X}^T \tilde{X})^{-1}, \text{cov}(\hat{f}) = \hat{\sigma}_e^2 S S^T \tag{15.9}$$

where $\hat{\sigma}_e^2$ is computed from the residuals, and any general smoothing procedure can be used. For the generalised least squares case these expressions become

$$\begin{aligned} \text{cov}(\hat{\beta}) &= (\tilde{X}^T V^{-1} \tilde{X})^{-1} \tilde{X}^T V^{-1} (I - S) V (I - S)^T V^{-1} \tilde{X} (\tilde{X}^T V^{-1} \tilde{X})^{-1} \\ \text{cov}(\hat{f}) &= S V S^T. \end{aligned}$$

Note that the intercept term is not included in X. Model (15.8) can be readily extended to handle more than one smooth function by incorporating the back-fitting algorithm described above. (See Speckman, 1988, for a discussion of 'partial smoothing splines'.)

For large samples the matrix S $(n \times n)$ may be too large to store. To avoid this problem, we define the following matrices. We assume that the number of distinct data points (r) is not too large (in large samples we typically will have several observations at many of the data points), so that storage is feasible. If not it may be possible to group data points or to interpolate. We define a matrix $Z(N \times r)$ where the k-th row has an element 1 in the position corresponding to the data point for the observation associated with the k-th row, and 0 elsewhere. The $(r \times r)$ matrix W has as the i-th row the set of weights associated with each data point for an observation located at the i-th data point. We can now write

$$S = ZWZ^T$$
$$SX = (ZW)(Z^T X)$$

which is the product of a $(N \times r)$ and a $(r \times p)$ matrix, where X is $(N \times p)$. We obtain a similar expression for Sy, so that (15.8) is readily evaluated. To ensure that the weights sum to 1, we form $A = diag(SJ) = diag((ZW)(Z^T J))$ where the diagonal contains the row sums of S, where J is a vector of ones. We then normalise S by writing $S^* = A^{-1}S = (A^{-1}Z)(WZ^T)$.

We note that this procedure has the advantage of requiring the storage of matrices of order $(N \times r)$ and $(r \times r)$ rather than $(N \times N)$. If there are data points that do not coincide with design points, then we need to compute the corresponding rows of SX and SY directly. An alternative is to interpolate using the nearest two design points, which involves two nonzero entries in the relevant rows of the matrix Z. This may often be an adequate approximation and is particularly useful in the case of a multilevel model where Su is recomputed at each iteration (see below).

Having estimated the parameters of the model, we may wish to compute the smoothed function over a new set of values of Z, including values not covered by the data points. In this case we, can calculate each row of SX and SY for each of these values, together with a corresponding value of the bandwidth, if this is variable, again obviating the need to store S.

The standard procedure for determining the smoothing parameter, typically the bandwidth, is based on cross validation. Several procedures are available (see, for example, Hastie and Tibshirani, 1990) which are based upon mean square error minimisation. A straightforward 'jacknife' procedure is as follows. For each point the data are smoothed leaving out that point, equivalent to setting the smoothing weight for the point to zero and rescaling the remainder to sum to 1. The cross validation sum of squares is then computed as

$$\sum_i (y_i - \hat{f}_\lambda^{-i}(x_i))^2 \tag{15.10}$$

where $\hat{f}_i^{-i}(x_i)$ is the fit at x_i computed by leaving out that point. The value of the smoothing parameter is chosen that minimises (15.10).

This procedure, however, assumes a constant bandwidth, which may be too restrictive, and we return to this later in Section 15.8.

15.5 Multilevel smoothing models

We first review some fully nonparametric models developed principally for longitudinal repeated measures data and then look at semiparametric models that allow full extensions to the general multilevel case.

Wu and Zhang (2002) propose a fully nonparametric procedure for longitudinal data, which in summary is as follows. For the j-th individual at time point i write

$$\begin{aligned} y(t_{ij}) &= f(t_{ij}) + v_j(t_{ij}) + e_{ij} \\ i &= 1, \ldots . n_j \end{aligned}$$

where the fixed part of the model $f(t_{ij})$ and the random component for individual j, $v_j(t_{ij})$, are both smooth functions. We choose a set of design points at which we shall evaluate the fixed and random functions. In some designs this set may coincide with the data points, especially where there are a few of these in common, but in general we will want to select a set spanning the whole data range for reporting purposes. At any design point, t, the fixed and random parts can be approximated using a Taylor expansion by local polynomials, for example, straight lines of the form

$$\begin{aligned} f(t_{ij}) &\approx f(t) + f'(t)(t_{ij} - t) \\ v_j(t_{ij}) &\approx v_j(t) + v'_j(t_{ij})(t_{ij} - t) \end{aligned} \tag{15.11}$$

Note that if the design points coincide with the data points we would use $f(t_{ij}) = f(t)$, $v_j(t_{ij}) = v_j(t)$, the 'local constant' estimator. Equation (15.11) can be written in the form

$$\begin{aligned} f(t_{ij}) &= X_{ij}\beta \\ v_j(t_{ij}) &= X_{ij}u_j \end{aligned} \tag{15.12}$$

At time point t we now minimise the 'local likelihood' based function

$$\begin{aligned} &\sum_j [y_j - X_j(\beta + u_j)]^T W [y_j - X_j(\beta + u_j)] + u_j^T \Omega_u^{-1} u_j \\ &\qquad + \log|\Omega_u| + \log|R_j| \\ &W = K_{jh}^{0.5} R_j^{-1} K_{jh}^{0.5} f \end{aligned} \tag{15.13}$$

where the kernel weighting function for the j-th individual is

$$K_{jh} = diag\{K_h(t_{j1} - t), \ldots . K_h(t_{jn_j} - t)\}$$

for bandwidth h where n_j is the number of observations for individual j. We can use any suitable smoother, such as the normal kernel. This is equivalent to the extended

maximum likelihood estimator, defined at t, given in Appendix 2.3. Note that the estimates are made separately at each design point and the smoothness is introduced by the kernel function, which is defined across the total set of design points for each individual. Estimation is straightforward and simply involves a weighted fit. To do this we define the level 1 explanatory variable for the random part as

$$K_{jh}^{-0.5} J, J = (1, \ldots .1)^T$$

and then fit the 2-level model in the usual way with the fixed and random explanatory variables defined as above. This provides estimates for the β, u_j, from which we can obtain the smoothed mean and individual curves using (15.12). Likewise, in the usual way, we can also obtain variance estimates for the fixed and random parameters, random effects and predictions at each design point, and we may wish to study how these change with time. To determine the optimum bandwidth, we can use a series of trials with visual inspection or a cross-validation procedure (see Wu and Zhang, 2002 for a discussion). We note that both the fixed and random parts are estimated nonparametrically.

A variant of this procedure is given by Wang (1998), who uses a non parametric kernel density smoother for the fixed part but for the random part considers parametric polynomial terms, as in the standard repeated measures model. The Verbyla *et al*. (1999) model is similar but using a penalised likelihood smoother and Carpenter *et al*. (2003) adapt this for discrete responses.

This model can be extended in several directions. If we have additional fixed effects and random effects, for example, at a higher level, we can write the likelihood in (15.13) as

$$\begin{aligned}\sum_j [y_j - X_j(\beta + u_j) - X^{(1)}\beta^{(1)} - X^{(2)}v_j^{(2)}]^T W[y_j - X_j(\beta + u_j) - X^{(1)}\beta^{(1)} \\ -X^{(2)}v_j^{(2)}] + u_j^T \Omega_u^{-1} u_j + \log|\Omega_u| + v_j^T \Omega_v^{-1} v_j + \log|\Omega_v| + \log|R_j|\end{aligned} \tag{15.14}$$

where $X^{(1)}\beta^{(1)}$ is the additional fixed part component and $X^{(2)}v_j^{(2)}$ the additional random effects associated with individual j. These additional components are also estimated nonparametrically, with the random effects and fixed and random parameters varying across design points.

One interesting application of this model is to the fitting of nonlinear growth curves to a sample of individuals (see Chapter 9) where the curves have lower or upper asymptotes or both and where the timing of the asymptotes varies across individuals. While the procedure does not provide actual asymptotes, a careful choice of bandwidth and polynomial order can be used provide a good approximation. Likewise, if we require monotonicity for the smooth curves, this can generally be achieved through a suitable choice of bandwith and polynomial order.

We now describe a general procedure to fit a semiparametric model that incorporates many of the previous models.

15.6 General multilevel semiparametric smoothing models

A general semiparametric random effects model can be written as

$$Y = X\beta + f(t) + Zu + e \tag{15.15}$$

As in Appendix 2.3, we have the extended likelihood function for model (15.15) in the form

$$\begin{aligned} 2L(\theta, \beta, u) = &- \log|R| - (Y - f(t) - X\beta - Zu)^T R^{-1} \\ &\times (Y - f(t) - X\beta - Zu) - \log|\Omega_u| - u^T \Omega_u^{-1} u \end{aligned} \tag{15.16}$$

We now use the procedure suggested in Appendix 2.3, iterating between (15.16), which describes a single level model and is used to calculate the fixed effects including the smoother, then calculating the random parameters as in the standard IGLS algorithm and then the random effects using the formulae in Appendix 2.2. Note that we can use any reasonable smoothing procedure.

Having calculated the fixed effects from (15.16), we form the fixed predictor and adjusted response

$$\begin{aligned} &X\beta^{(1)} + f^{(1)}(t) \\ &\tilde{Y} = Y - (X\beta^{(1)} + f^{(1)}(t)) \end{aligned} \tag{15.17}$$

and then compute updated estimates of all the random parameters and hence the random effects using

$$\tilde{Y} = Zu + e.$$

With these estimates we compute new estimates of the fixed part of the model and iterate until convergence. In this formulation all the random effects are assumed to be parametric. Inference can be based upon the large sample covariance matrix estimates. Large sample covariance estimates are given by the generalised least squares version of (15.9), with V derived from the random parameter estimates. The marginal estimate of f is given by

$$\hat{f} = S(y - X\hat{\beta}).$$

When computing S, we an use the same procedure for forming S, as above. Now, however, Su is required and is recomputed at each iteration, unlike SX and SY which need only be computed once. This suggests that there may be a considerable computational advantage in ensuring that the design points correspond with the data points.

Hu *et al*. (2004) discuss an approach with similarities to the present one, but it is based upon generalised estimating equations (GEE) and so is not a true multilevel model, and does not provide estimates for the random parameters.

When choosing the bandwidth based upon a cross-validation (CV) criterion, we will track the fixed part of the model. Corresponding to the least squares criterion, we have a generalised least squares criterion

$$(Y - X\beta - \hat{f}_\lambda^{-})^T V^{-1} (Y - X\beta - \hat{f}_\lambda^{-}) \tag{15.18}$$

where $\hat{f}_\lambda^{-}$ is the vector formed by the jacknife procedure described above, omitting each design point in turn. Because of the iterative estimation procedure, the calculation of (15.18) will be time-consuming; a modification would be to use a common estimate of V derived from an initial fit which ignores the random effects, or one which uses an approximate parametric model.

For the fixed part of the model, we can compute the prediction variance using (15.5); for the random parameters and random effects we can use the usual procedures for that part of the model. We recall that the procedures for nonparametric repeated measures models can be defined in terms of a standard random coefficient model, with explanatory variables defined by the kernel weighting functions. It follows, therefore, that the procedure of the present section is readily adapted to include nonparametric (repeated measures) components and we can fit any mixture of parametric and nonparametric components in the fixed and random parts of the model.

15.7 Generalised linear models

We can adapt our procedures to the case of multilevel generalised linear models with discrete responses. Without going into details, essentially the nonlinear model is linearised using MQL or PQL procedures and then the above iterative procedure can be used. In the linearisation, we compute first and second differentials for the link function and apply these to the additive combination of the parametric predictors and the smoothed predictor(s), leading to quasilikelihood estimators. Thus, at the t-th iteration we replace $X_{ij}\hat{\beta}_t$ by $X_{ij}\hat{\beta}_t + \hat{f}_t$, etc. We can also obtain marginal likelihood estimators using, for example, quadrature or simulated maximum likelihood (Ng *et al*., 2006).

15.8 An example

For comparison with a simple polynomial smoother, we fit a 2-level variance components model in MlwiN (Rasbash *et al*., 2008) where we obtain the estimates in Table 15.1 for a quadratic model where class is level 2 and student is level 1.

The data are a subset of those used by Blatchford *et al*., (2002) using class sizes between 18 and 33, centred on 30 with a normalised mathematics test score as outcome. There are 4829 pupils in the present analysis and two covariate pre-test scores for maths and literacy are included.

Table 15.1 Post reception mathematics score by class size. Variance components model.

	Estimate	Standard error
Fixed		
Intercept	−0.95	
Class size	−0.044	0.014
Class size^2	0.003	0.002
Pre-test Mathematics score	0.407	0.015
Pre-test literacy score	0.019	0.001
Random		
Between-class variance	0.233	0.023
Between-pupil variance	0.393	0.008

Class size is the regular class size averaged over all three terms that a child was in the school.

The class size function plot, estimated at the mean values of the covariates, is given in Figure 15.1

Figure 15.2 shows the 95 % confidence intervals for each data point.

We fit the semiparametric (multilevel partial spline) model described in Section 15.3 using a normal kernel density estimator, and the procedure for handling large smoothing matrices. To illustrate the effect of choosing different bandwidths we show, in Figure 15.3, the plots and parameter estimates for a narrow bandwidth of 0.45 (quartic kernel), for one of 2 (Normal kernel) and one of 4 (quartic kernel) with increasing degrees of smoothness from a curve that closely follows each data point

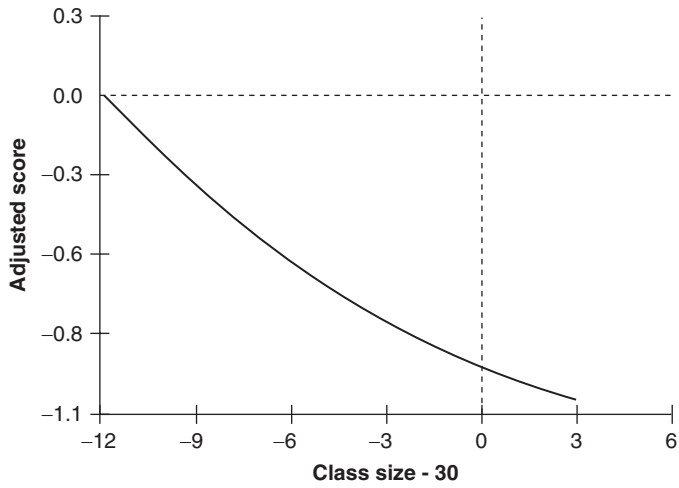


Figure 15.1 Two level model quadratic class size plot.

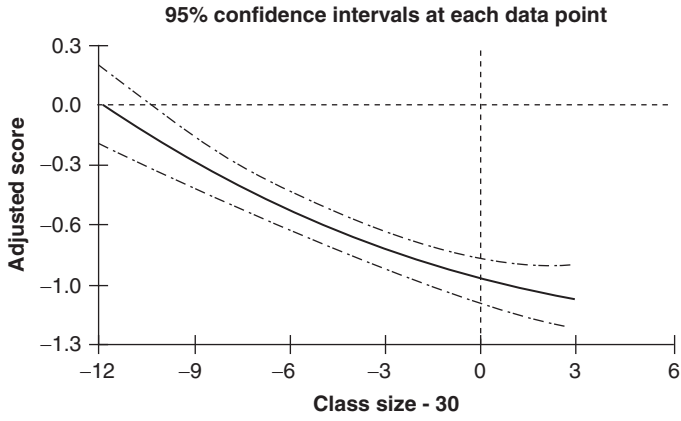


Figure 15.2 Confidenc intervals for predicted response.

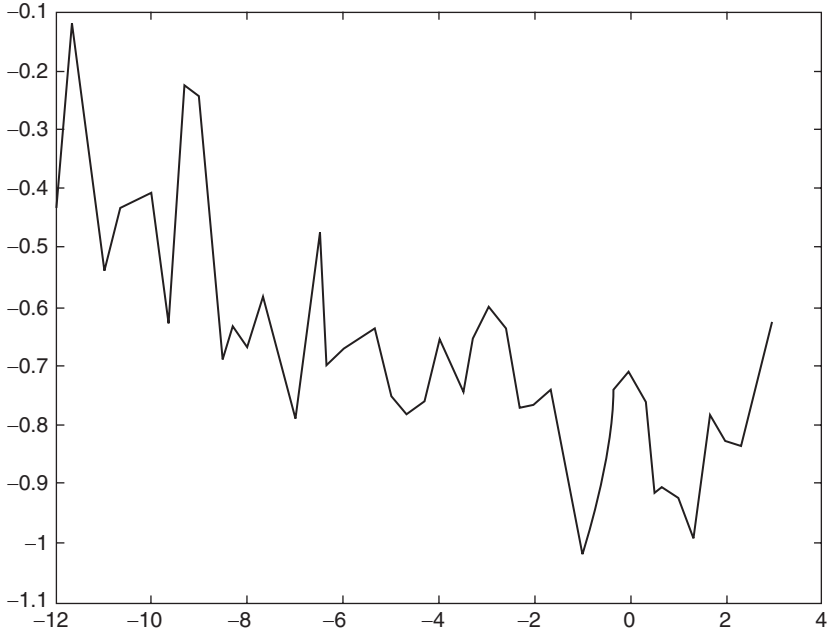
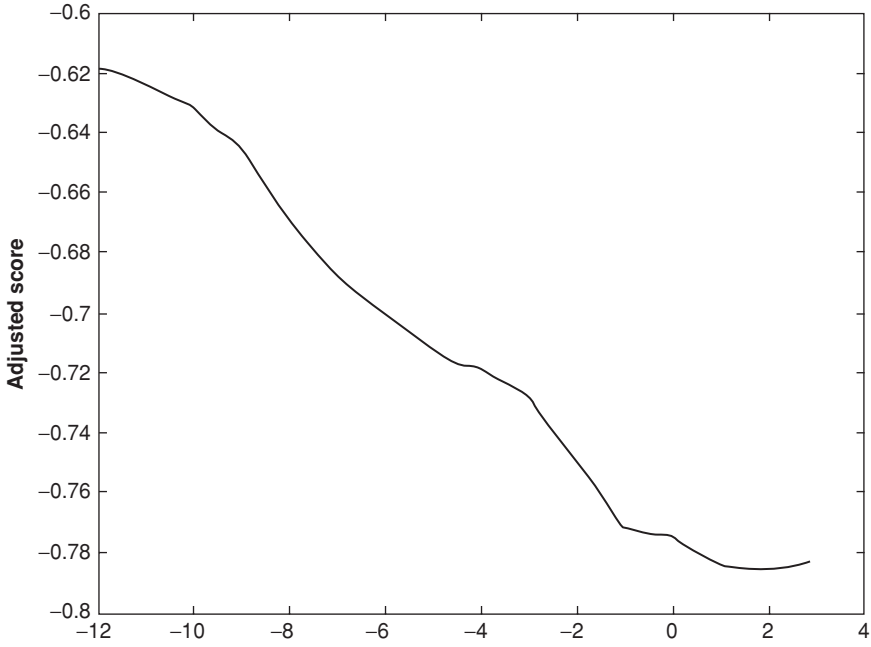
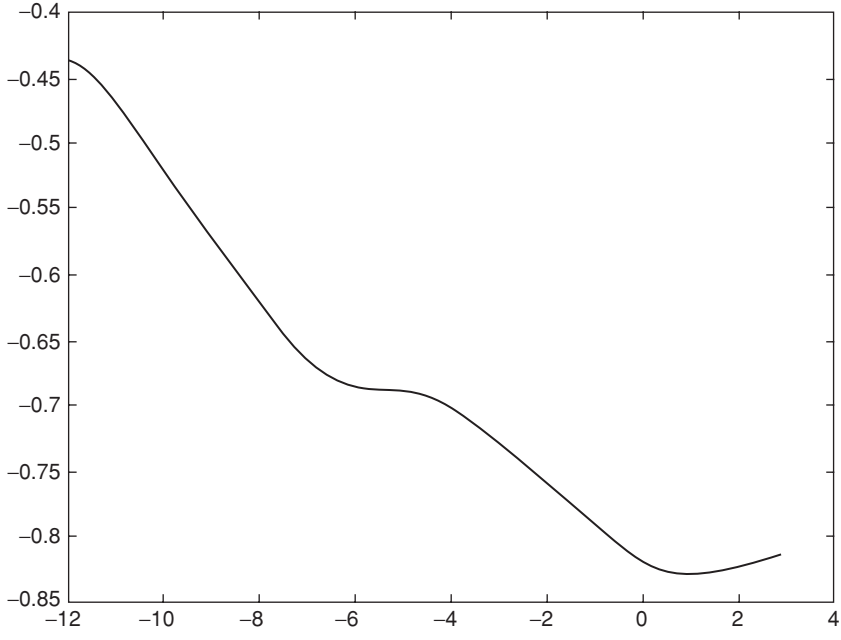


Figure 15.3 Smoothed class size relationships for different bandwidths and kernels. 1. Bandwidth = 0.45; quartic kernel. Horizontal axis class size measured about 30.



2. Bandwidth = 2; normal kernel



3. Bandwidth = 2; 4; quartic kernel

Figure 15.3 (Continued)

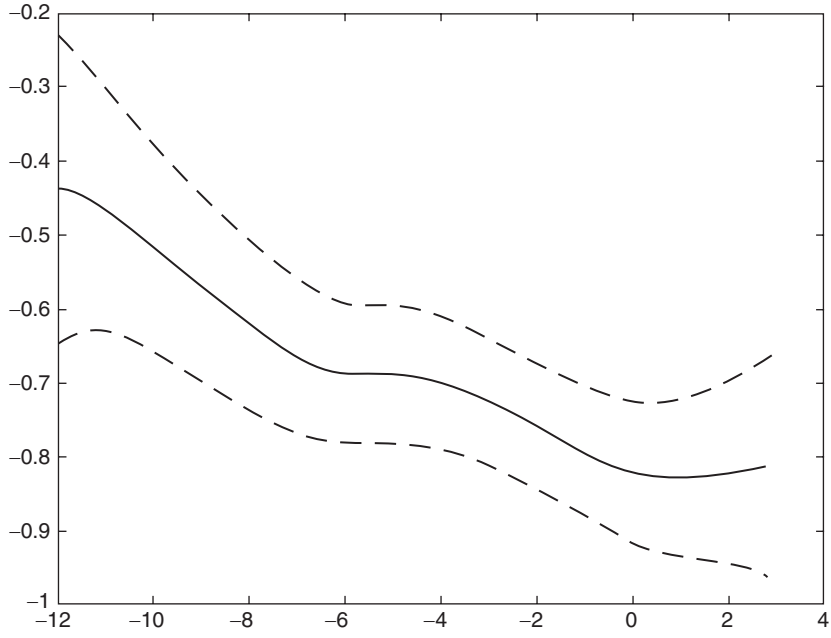


Figure 15.4 95% confidenc intervals for fitte smoothed line. Bandwidth = 4; quartic kernel.

to one where there is a fairly smooth curve. The numbers of iterations required were respectively 17, 35 and 12.

Figure 15.4 shows the 95% normal confidence interval for the fitted line with bandwidth = 4; quartic kernel.

The fixed part estimates for pre-test mathematics and literacy scores were similar for all bandwidths (0.405 (0.015) and 0.019 (0.001) for the bandwidth of 4 with a quartic kernel) as were the variance estimates (0.211 (0.021) and 0.397 (0.008) for a bandwidth of 4 with a quartic kernel). The fixed part estimates are very close to those for the parametric model, although the level 2 variance is a little different.

The first plot in Figure 15.2 has a small bandwidth so that the plot follows the actual data closely and shows a rather discontinuous relationship, albeit with an overall trend. Comparing this with Figure 15.1, we see a similar range for the adjusted response with the quadratic estimating a value at the smallest class size somewhat beyond the range of the data. For the bandwidth of 2 using a normal kernel we have a relatively smooth relationship with asymptotes at either end of the class size scale, with a difference in adjusted score between the smallest and largest classes of 0.16, compared to a difference of 0.7 for the quadratic fit. For the quartic kernel with a bandwidth of 4 the corresponding difference is 0.40. With increasing bandwidth the difference decreases, and somewhat more markedly for the normal kernel.

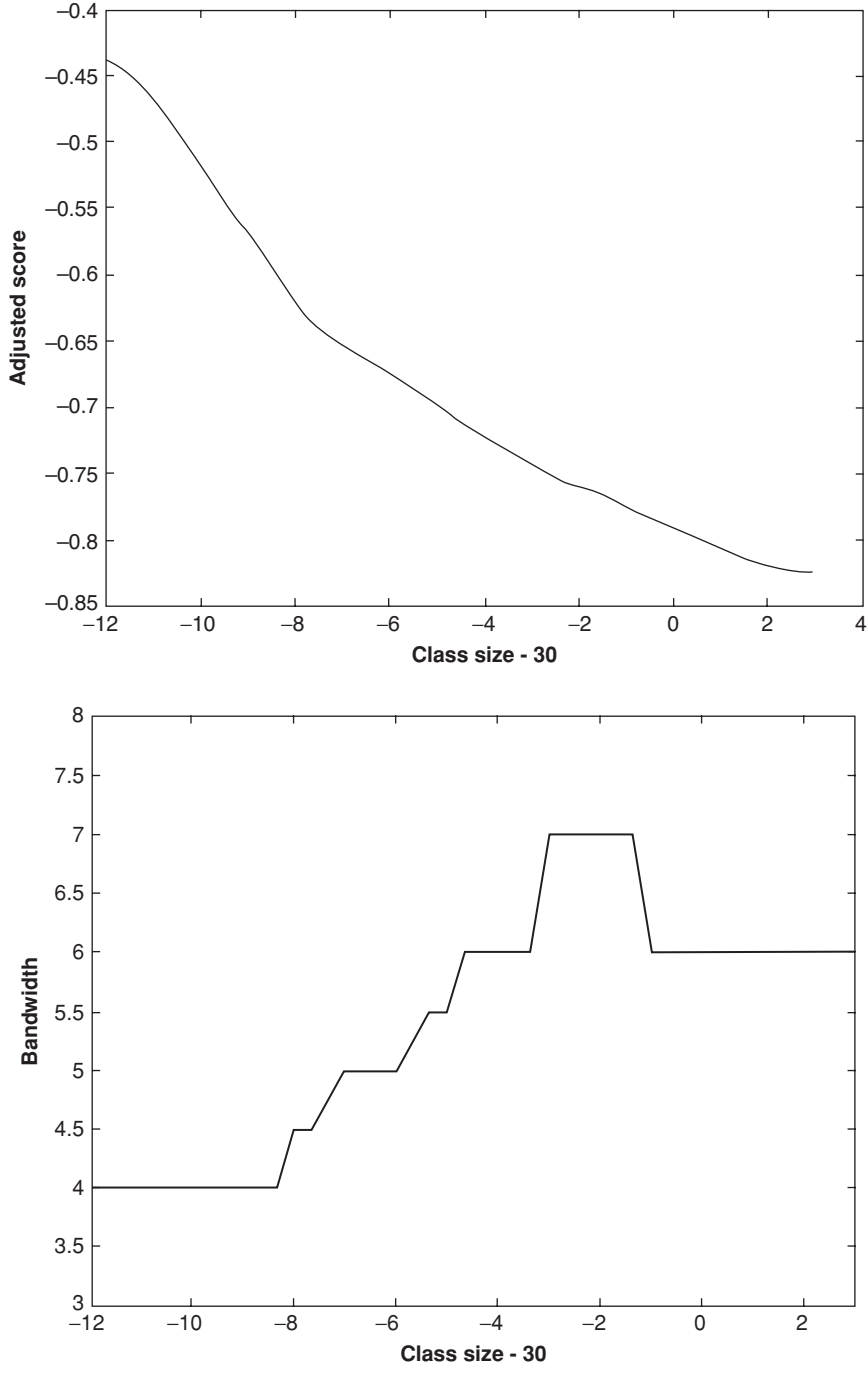


Figure 15.5 Fitted smooth line with variable bandwidth.

So far we have considered a constant bandwidth. This restriction is unnecessary, however, and the bandwidth can vary across design points as pointed out earlier, and the computations are easily adapted. We have adjusted the bandwidth across the design points in order to eliminate, for the bandwidth of 4, the flattening of the curve for class sizes of 24–26 and the upturn for class sizes of 31 and above. The resulting smooth curve is plotted in Figure 15.5, which also displays the values of bandwidth used.

15.9 Conclusions

We have shown how a general multilevel semiparametric smoothing model can be specified and fitted using a straightforward iterative procedure. One problem with parametric regression spline models is that the choice of join points can be difficult, whereas in the former case the only important choice is that of bandwidth.

The example illustrates the importance of choosing an appropriate bandwidth; in the present case a value of about 2 for the normal and 4 for the quartic appear appropriate, since they produce a relatively smooth curve that is (almost) monotonically related to class size, which seems to be a reasonable requirement for these data. It also illustrates the complexity of bandwidth choice when we allow the bandwidth to vary across the range of the design points. This suggests that the use of automatic cross-validation techniques to choose bandwidth may be infeasible when bandwidth is allowed to be variable. Rather, choice should be guided by what may be assumed to be reasonable in the light of the actual pattern in data being fitted and assumptions about what might be considered legitimate. The presentation of confidence intervals is also important. The example also illustrates a particular problem with the simple, quadratic, parametric model where end-point extrapolation seems to produce inflated estimates.

Because the proposed algorithm for the semiparametric model requires the estimation of the random parameters and effects conditionally on the fixed part of the model at each iteration, we can incorporate complex random and hierarchical structures. These will include, for example, cross classifications and complex variation at level 1.

Finally, we can fit any combination of parametric and nonparametric components in both the fixed and random parts of the model. The one limitation is that if we fit a nonparametric random effect and a parametric random effect at the same level we assume independence within the present framework. This may not always be realistic, however, and further work is needed into ways of allowing for patterned dependencies across the nonparametric design points.

16

Missing data, partially observed data and multiple imputation

16.1 Introduction

A characteristic of many studies is that some of the intended measurements are unavailable. In surveys, for example, this may occur through chance or because certain questions are unanswered by particular groups of respondents. An important distinction is made between situations where the existence of a missing data item can be considered a random event and those where it is informative, and the result of a nonrandom mechanism. Randomly missing data may be missing 'completely at random' (MCAR) or 'at random' (MAR) conditionally on the values of other measurements. Otherwise it will be 'missing not at random' (MNAR). Where data are MNAR we have a choice of methods, all of which rely on making particular assumptions about the mechanisms creating missingness. Some of these rely upon strong distributional assumptions (see, for example, Heckman, 1979), others focus on making use of 'auxiliary' data that are associated with the probability that a data item or record is missing and are also associated with the variables that contain the missing values. In practice, deciding how and why values are missing is not always straightforward and we discuss below the kinds of assumptions that need to be made.

Where missingness applies to whole records in a dataset, one approach is to use auxiliary data, along with variables in the model of interest (MOI), to set out a model that predicts the probability that a record is missing. We discuss such a model in Section 16.3. One drawback of this procedure is that it does not handle cases where individual items or variables in a record are missing, and we shall not consider it in

Multilevel Statistical Models: 4th Edition Harvey Goldstein

any detail. Instead, we focus on multiple imputation procedures that can handle quite general patterns and types of missingness.

The statistical tools that we introduce can handle not only cases where intended measurements are unavailable, but also a wide range of problems from attrition in longitudinal studies, through situations where data values are known only approximately, to ways of efficiently handling record linkage data. We also examine the 'default' procedure that simply omits any record with missing data – the 'listwise deletion' or 'complete case analysis' method.

We consider the problem in two parts. First, we look at how to fill in or 'impute' the missing values. Once we have a 'completed' dataset we can then fit our model of interest in the usual way; we discuss below has this is handled. Detailed discussions of missing data procedures are given by Rubin (1987) and Little (1992). A useful introduction to ways of handling missing data, software and with examples can be found at http://missingdata.org.uk/. We begin by considering a single level model for simplicity.

In the process of describing how to deal with multilevel missing data, we shall draw upon the discussion of responses at several levels in Section 6.7 and also the latent normal model that was developed in Chapter 7.

16.2 Creating a completed dataset

Consider the standard single level linear model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i \tag{16.1}$$

for the i-th unit in a single level model. Suppose that some of the x_{1i} are MCAR or MAR conditional on x_2. Label these unknown values x^*_{1i}. We consider the estimation of these by predicting them from the remaining observations and the parameter set θ for the prediction model, namely

$$\hat{x}^*_{1i} = E(x^*_{1i}|x_{2i}, y_i, \theta) \tag{16.2}$$

Where we have multivariate normal data, the prediction model is simply the linear regression of X_1 on X_2, Y. We shall consider the case of non-normal data later.

We can define the following multivariate model with three response variables, Y, X_1, X_2 each related to an intercept

$$\begin{pmatrix} Y \\ X_1 \\ X_2 \end{pmatrix} \sim N(\mu, \Omega), \mu^T = (\mu_Y, \mu_{X_1}, \mu_{X_2}), \quad \Omega = \begin{pmatrix} \sigma^2_Y & & \\ \sigma_{YX_1} & \sigma^2_{X_1} & \\ \sigma_{YX_2} & \sigma_{X_1X_2} & \sigma^2_{X_2} \end{pmatrix} \tag{16.3}$$

The parameters of this model, where responses are MCAR can be estimated using the procedures described in Appendix 6.1. We can then form the conditional

distribution

$$X_1|X_2, Y \sim N(\mu, \sigma^2), \quad \mu = \mu_{X_1} + \beta_Y(Y - \mu_Y) + \beta_{X_2}(X_2 - \mu_{X_2}),$$
$$\begin{pmatrix} \beta_Y \\ \beta_{X_2} \end{pmatrix} = \begin{pmatrix} \sigma_Y^2 & \\ \sigma_{YX_2} & \sigma_{X_2}^2 \end{pmatrix}^{-1} \begin{pmatrix} \sigma_{YX_1} \\ \sigma_{X_1X_2} \end{pmatrix}, \quad \sigma^2 = \sigma_{X_1}^2 - \sigma_{X_1}^{-2} \begin{pmatrix} \sigma_{YX_1} \\ \sigma_{X_1X_2} \end{pmatrix}^T \begin{pmatrix} \sigma_{YX_1} \\ \sigma_{X_1X_2} \end{pmatrix} \tag{16.4}$$

The multiple imputation procedure then samples sets of missing values from this distribution, including the uncertainty involved in *estimating* the parameters in (16.4). Before discussing this in detail, we now look at some alternative possible ways of imputing missing values. Perhaps the simplest procedure for dealing with missing data is simply to remove any records containing missing values from the dataset. As long as the missingness mechanism is independent of the response variable, *Y,* we will obtain unbiased estimates. This 'listwise deletion' procedure, however, is inefficient since it results in records being lost. It also cannot deal with cases where the missingness is informative. Other methods are 'single' imputation methods that seek to replace each missing value by a single value and then fit the MOI.

A simple single imputation scheme is to compute the mean value for each variable using the nonmissing values and then to use this to fill in the missing values. This procedure will tend to underestimate the variance for such a variable and thus lead to biased parameter estimates. Likewise, using (16.4) to obtain a single value of X_1 predicted using the first line of (16.4) will also tend to underestimate the variance of X_1. So-called 'hot deck' methods are often used in large scale surveys. These are nonparametric methods where a data record containing missing values is 'matched', using a subset of fully observed matching variables, with a set of similar records elsewhere in the data file that are complete, and then one of these is randomly selected. While this method can sometimes be treated in similar ways to the MI methods under discussion here, a common problem is that it may not be possible to find a good match without a great deal of 'coarsening' the data categories or groups (Rubin, 1987), depending on which matching variables are chosen; this is especially so in multilevel data where matching should be within the same higher level unit, or where higher level units should also be involved in the matching criteria. In all the single imputation methods, because only a single completed dataset is used, the full uncertainty associated with the missing data values is not carried through to the analysis and thus will lead to biased standard error estimates.

Another method uses weights that are derived from a model relating the probability of a record being missing to auxiliary variables and any fully observed variables in the MOI. The inverses of the predicted probabilities from such an analysis are then used as weights in an analysis using the fully observed cases. The problem with this approach is that it will generally lose efficiency, although methods to recover some of this have been developed. Thus, Carpenter *et al*. (2006) discuss 'doubly robust' methods that yield unbiased estimates if *either* the imputation method is correct *or* if the weights are correct. These methods do not deal very well with the case when there is a general pattern of missing data, as opposed to the simpler case when only a small number of missing data patterns exist. They also rely upon good estimates of the weights, but since these are estimated from the data they may not be very reliable.

16.3 Joint modelling for missing data

Where missingness is informative, that is we have MNAR, it may be useful to adopt a fully model based procedure. For the case where we have missing records, we need to write down an explicit model for the probability of being missing and link this to the model of interest that we wish to fit. Consider a simple situation where the MOI is given by (16.5) and the response probability model is given by (16.6).

$$y_{1i} = \beta_0 + \beta_1 x_i + e_{1i} \tag{16.5}$$

$$\text{Pr(response observed)} = \int_{-\infty}^{\alpha_0+\alpha_1 z_i} \phi(t)dt \tag{16.6}$$

where the Z is an auxiliary variable (or more generally a set of variables) that predicts the probability of a response (Y_1) being observed and for which there is the underlying standard latent normal variable $y_{2i} = \alpha_0 + \alpha_1 z_i + e_{2i}$. We assume that Z is uncorrelated with the assumed bivariate normal random effects e_{1i}, e_{2i}. The two models are linked by assuming

$$E(e_{1i}e_{2i}) = \sigma_{e12} \neq 0 \tag{16.7}$$

The existence of $\sigma_{e12} \neq 0$ is a statement that the propensity for a record to be missing is related to the response in the MOI, that is MNAR, after adjusting for covariates. In such a case a joint analysis will yield unbiased estimates. This is conveniently carried out in a Bayesian framework where the missing values are treated as parameters to be estimated and samples from the appropriate posterior distributions are chosen at each iteration of an MCMC algorithm. Alternatively, Heckman (1979) suggested a two-step procedure that involves a separate estimation of the covariance matrix, but this will generally not be as efficient as a full joint estimation. We note that if we have different categories of missingness, the above model can be extended using a latent normal formulation for a multicategory response as set out in Chapter 7, yielding an underlying multivariate normal distribution rather than just a bivariate normal.

While such a full model based procedure is often attractive, it does not deal with the case where individual items in a record additionally may be missing. In order to extend this model to the case where there is item missing data, we can use MCMC, assuming MAR for the missing items, and treating all the missing values as additional parameters, but this becomes computationally very time consuming. Further references are those of Nathan (1983), Nathan and Holt (1980) and Pfeffermann (2001). The multiple imputation procedure described below will handle the case where whole records are missing as well as missing items. Before we describe this, we bring together Appendix 6.1 and Chapter 7 to show how to impute missing responses in a multilevel model where the responses can be at any level and of mixed type. We illustrate with a 2-level model.

16.4 A 2-level model with responses of different types at both levels

In Chapter 6, we set out a 2-level multivariate model with responses at both levels as follows

$$\begin{aligned} y_{ij}^{(1)} &= X_{1ij}\beta^{(1)} + Z_{1ij}u_j^{(1)} + e_{ij}^{(1)} \\ y_j^{(2)} &= X_{2j}\beta^{(2)} + Z_{2j}u_j^{(2)} \\ e_{ij}^{(1)} &\sim MVN(0, \Omega_1), u_j = \left(u_j^{(1)}, u_j^{(2)}\right)^T, u_j \sim MVN(0, \Omega_2) \end{aligned} \tag{16.8}$$

where the superscripts indicate the level. We can add further classifications such as a cross classification as described in Section 12.6.

In Chapter 7, we showed how to sample a multivariate latent normal given categorical responses and we now set out the steps needed to combine these several types where we have several levels or classifications. We shall assume that there are two levels but the steps are readily extended to more than two classifications.

16.4.1 Sampling level 1 non-normal responses

When sampling a non-normal response at level 1 we condition on the fixed part of the model associated with the response, together with any latent normal or observed normal variables at the same level that are correlated with the response. Thus, for example, in the binary case we can write the conditional distribution as in (7.2) as

$$\begin{aligned} &y_{ij,1}|y_{ij}^* \sim N(X_{ij}\beta_1 + y_{ij}^*\beta_2 + z_{ij_1}u_{j_1} + z_{ij_2}u_{j_2} + \ldots z_{ij_q}u_{j_q}, 1 - \Sigma_{21}\Sigma_1^{-1}\Sigma_{12}) \\ &\beta_2 = \Sigma_{12}\Sigma_1^{-1}, \Sigma_1 = \text{cov}(y_{ij}^*), \Sigma_{21} = \text{cov}(y_{ij.1}, y_{ij}^*), j = \{j_1, j_2, ..j_q\} \end{aligned} \tag{16.9}$$

The subscript 1 refers to the variable being sampled and 2 refers to the correlated level 1 variables. The subscripts $j_1, j_2, ..j_q$ refer to the classifications, which might be further hierarchical levels, cross classifications or multiple membership structures. The * denotes the variables being conditioned on. Using the steps given in Chapter 7, we can also then sample the level 1 fixed coefficients, and the level 1 residuals by subtraction. The level 1 covariance matrix will have variance elements constrained to 1 corresponding to categorical responses and covariance elements set to zero for unordered multicategorical responses, and the steps set out in Chapter 7 can be used. In the case of the level 2 residuals, we sample for each classification in turn, conditioning on the remaining residual estimates. For binary, ordered or continuous non-normal responses, we sample a single latent normal variable and for a *p*-category unordered variable we sample *p-1* latent normal variables.

16.4.2 Sampling level 2 non-normal responses

We proceed to sample these in a similar fashion to those at level 1 where the level 2 model is given by the second line of (16.8). For simplicity, we assume a simple residual at level 2 but we can extend the sampling to include the complex variance case.

The level 2 residuals for level 2 responses are obtained by subtraction, but the level 2 covariance matrix now includes the level 2 responses as well as the level 2 residuals for the level 1 responses. This is therefore sampled in a single step, noting that some elements will be constrained to 1 or 0 if we have categorical responses at level 2.

Each classification will be sampled in turn, conditioning on the remaining classification residual estimates. If a level 1 or level 2 response is missing then we sample a value from the latent multivariate normal distribution as described in Appendix 6.1. For such missing values we can transform these back to the original non-normal scale using the current parameter values. For example if the original response was binary then we choose the value to be 1 if it exceeds the current threshold parameter value, etc.

16.5 Multiple imputation

The usual multiple imputation procedure (Rubin, 1987) works as follows. The predicted values are adjusted to have their correct, on average, distributional properties by sampling from the posterior (predicted) distribution of all the variables with missing values. Using MCMC preserves the full uncertainty associated with the estimation and is a convenient way to implement the MI procedure. It also allows the use of informative priors.

Having generated a 'completed' dataset with all missing values imputed, we then fit our multilevel model in the usual way and obtain parameter estimates. This process is repeated m times and the final estimates are suitably chosen averages of these sets of estimates.

For a parameter θ with point estimates θ_i and variance estimates $\text{var}(\theta_i)$, we form

$$\begin{aligned} \hat{\theta} &= \sum_{i=1}^{m} \theta_i / m, \quad \text{var}(\hat{\theta}) = \bar{V} + (1 + m^{-1})B \\ \bar{V} &= \sum_{i=1}^{m} \text{var}(\theta_i)/m, \quad B = \sum_{i=1}^{m} (\theta_i - \hat{\theta})^2/(m-1) \end{aligned} \tag{16.10}$$

with corresponding expressions for a set of parameters and their covariance matrix. The value of m has to be chosen and for multilevel data a minimum value of 10 is generally advisable. Assuming that these point estimators are asymptotically normal, these final estimates are efficient with consistent standard errors.

To implement this procedure, we choose all those variables in the MOI with missing values to be responses in a multivariate model. This may include variables at any level and normal as well as categorical variables. We then fit this model as described in the previous section and where values are missing these are imputed

as described, on their original scales. The variables with no missing values can be used as explanatory variables in this imputation model or incorporated as further responses. In carrying out MI, it is important that all the variables in the MOI are used for imputation and also where the MOI contains a multilevel structure, that this is incorporated also (Goldstein *et al*., 2009, give examples).

Where we have auxiliary variables we can use these as covariates in the imputation model and if these are associated with the missingness mechanism, conditioning on them can reduce any biases by increasing the plausibility of the MAR assumption. Thus, we see that MI readily incorporates procedures for dealing with informatively missing data. We can also incorporate the case where whole records are missing since in that case we will impute all the missing variables, again conditioning on the auxiliary variables. In a later section we shall return to this when discussing longitudinal data. Software (REALCOM, 2008) has been developed to carry out this procedure.

16.6 A simulation example of multiple imputation for missing data

We use a simulation described in Goldstein *et al*. (2009). The data set is the 'tutorial' dataset, distributed as an example with the *MLwiN* software package (Rasbash *et al*., 2008). This consists of a normalised measure of educational achievement at 16 years (the response) and a number of covariates at both level 1 (student) and level 2 (school), as detailed in Table 16.1.

The following MOI is first fitted to the complete dataset:

$$y_{ij} = X_{ij}\beta + u_j + e_{ij}$$
$$e_{ij} \sim N(0, \sigma_e^2), u_j \sim N(0, \sigma_u^2)$$

This gave the parameter estimates in the second column of Table 16.1. Of the 65 schools, 10 were randomly sampled for each simulated data set where each school had the same probability of inclusion. For these the school gender was set to be missing. In addition, 10 % of the response values were randomly set to be missing and 5 % of the verbal reasoning band categories were set to be missing. This yielded between 25 % and 30 % of records with any missing data. One hundred datasets were simulated.

With the exception of the school gender category 3, all the full data values lie within 95 % normal confidence intervals derived from combining the imputed datasets. The results show negligible biases for all the level 1 parameters and the variance estimates. For the level 2 categorical variable, school gender, there appears to be a small downward bias. There are two possible reasons for this. First, it may be due to the rather small number of schools in the study combined with the assumption of a uniform prior for the level 2 covariance matrix in the imputation model. Second, and more subtly, the MOI is not exactly compatible (congenial) with the joint distribution implied by the imputation model, although it is likely to be close because the

Table 16.1 Simulation study model. Parameter estimates and standard errors in brackets. One hundred simulated datasets. MCMC estimation used a burn in of 2000 with five imputed data sets at iterations 1, 500, 1000, 1500, 2000. Estimates are computed using restricted maximum likelihood.

Parameter	Complete dataset	Imputation	Relative bias (%)	Imputation standard error***
Intercept	0.260 (0.056)	0.263	1.2	0.0021
Reading test score	0.391 (0.017)	0.391	0.0	0.0007
Verbal reasoning band 2*	−0.417 (0.032)	−0.414	−0.7	0.0014
Verbal reasoning band 3*	−0.765 (0.054)	−0.768	0.4	0.0024
School gender category 2**	0.099 (0.108)	0.091	−8.1	0.0040
School gender category 3**	0.241 (0.084)	0.230	−4.6	0.0038
Level 2 variance	0.079 (0.016)	0.080	1.3	0.0005
Level 1 variance	0.536 (0.012)	0.536	0.0	0.0004

*Verbal reasoning band has three categories: category 1 (the reference category) is the top 25 % of original verbal reasoning scores, category 2 is the middle 50 % of verbal reasoning scores and category 3 is the bottom 25 % of verbal reasoning scores. **School gender has three categories: mixed schools (the reference category); category 2 is boys' school; and category 3 is girls' school. ***The imputation standard error is the standard error for each parameter over the 100 simulations.

imputation model relies on an underlying joint multivariate normal distribution with unstructured covariance matrices at level 1 and level 2 (see Goldstein *et al*., 2009).

16.7 Longitudinal data with attrition

A special case of missing data occurs in longitudinal studies where, at any given follow-up occasion or 'sweep' of the participants, some whole or partial records may be unavailable; for example, because of refusals or other reasons that may be related to the variables in the MOI. We may regard such missing records as an example of missing data where there are other variables, measured, for example, at the time of first contact or as auxiliary data collected by interviewers about the participant. Thus, in principle, we can apply the methods we have described that will also deal with individual items or questions with missing values.

This procedure has several advantages over traditional ones that depend on weighting to adjust for attrition. The methodology for computing weights specifically in order to compensate for 'informative' attrition that leads to biases, involves procedures for estimating the probability of a survey sample member responding, for each sample member at each occasion, as a function of sample member characteristics. Hawkes and Plewis (2006) provide a useful description of a model based approach to this. The resulting (inverse) probabilities are then used in standard ways

in subsequent analyses. One of the problems with such procedures is that response may be partial. Thus, for example, all individuals may respond to a set of educational variables but not to health variables at a particular occasion. In this case, we would not use weights to adjust for attrition when analysing the educational data, but we would wish to do so if health variables were additionally being analysed. This may complicate matters, because successive analyses would be based upon different numbers of individual cases. A further problem is that for a particular individual the weights will generally change from occasion to occasion, depending on the possibly changing response patterns across the sample, which may alter the joint distribution of respondent characteristics, and this will create difficulties when conducting analyses across several occasions.

Another issue when using weights is that a traditional weighting approach will generally lose data information. Suppose we wish to regress a time 2 variable on a set of time 1 variables with attrition occurring. Weights can be obtained in various ways but the weighted analysis will then be carried out using only the cases with measurements at both occasions. This ignores the information, available at one occasion but not at another, that is not incorporated in weights but may be available, for example, from related auxiliary variables or further covariates in a model. A similar problem arises in the case mentioned above where there is a differential response to, say, health and educational variables.

In some cases, we may have a basic set of weights derived from a complex sampling design, or computed to adjust for initial nonresponse, that we would wish to incorporate in a general MI analysis. Here, we would use weighted MCMC estimation as described in Section 3.4.2 for both the imputation model and the MOI, or in a weighted likelihood analysis if this is used for the MOI (see Goldstein, 2009, for further discussion).

16.8 Partially known data values

We may have a response where some values from a continuous distribution are known accurately but others are only known to lie within a given range. One example is retrospective data, where the time since an event is measured and where some individuals can only provide an interval estimate. An illustration is in the measurement of pregnancy gestation length where only some individuals can supply accurate values of the timing of their last menstrual period. Since, typically, we do not know which are the accurate values, an additional step is required to provide a probability distribution which will assign a probability for the observed value close to 1 where this value is in fact accurate, and a more variable distribution of values where it is not. Mixture modelling (see Chapter 8) provides one approach to this. Another, extreme, example is that of truncation where all measurements below a given value are known but for the remainder we only have the information that they lie above the given value. An extension of this is where we have several possible intervals and associated information about which interval an observation lies in. In other cases, we may have probabilistic information associated with different possible ranges of values for a

variable. For categorical responses, we may be able to specify a set of probabilities across a set of categories.

All of these are examples of data coarsening (Heitjan and Rubin, 1991), where what is observed is intermediate between fully missing and fully known. If we have probabilities associated with the possible data values, then we need a procedure for combining this probability distribution with the conditional distribution from the standard imputation step. We assume that the specified probability distribution is independent of the variables involved in deriving the conditional distribution for the missing value. We propose the following procedure.

Where we have a categorical variable, the standard imputation step yields an imputed (back-transformed) value, x, for a particular variable, X, for a particular data record. We select a value at random from the specified probability distribution for X, and if this is equal to x then this value is accepted. Otherwise, the imputation and the selection are repeated and this continues until a value is accepted.

For continuous variables, we propose that we can discretise the probability distribution into a set of ordered categories. The standard imputation step, as before, yields a normal value or a back-transformed value for a non-normal continuous variable, x for a particular individual; we select a category at random from the discretised distribution and if x belongs to this category it is accepted. Otherwise, a new imputation and selection are carried out until acceptance occurs. When using this procedure, we need to have sufficient categories to provide a good approximation.

This procedure is computationally equivalent to treating the posterior distribution associated with the standard imputation step as a likelihood for the unknown value of X and the specified probability distribution as a prior for it. For convenience, we shall therefore refer to the procedure as 'prior informed imputation' and in this context we refer, somewhat loosely, to the specified probabilities as constituting a 'prior' distribution.

Note that if more than one response for a unit has a partially known value, all responses have to be selected at one (multivariate) draw. As an example, for ordered categories, when computing the loglikelihood contribution in MH sampling, we form a weighted loglikelihood over the valid categories with the weights being the 'prior' probabilities.

An example for categorical variables is where individuals are asked to choose a response category but some cannot make as fine a distinction as other individuals, so that for these we can assign a prior over several categories. Another case is where data are to be coded, for example, into social groupings, but where for some individuals we only have an assignment to a group of several categories. We assume that the occurrence of such assignments is independent of the model parameter values.

We now look at an application of this procedure to probabilistic record linkage (see, for example, Scheuren and Winkler, 1993).

16.8.1 An application to record linkage

The linking of records from disparate sources is becoming increasingly important with the advent of large, typically administrative, databases and the facility to link

individual records across these and with surveys. Because of the inherent uncertainties associated with the matching of individuals, probabilistic matching is often used and the following description captures the essence of the procedure.

Consider a primary file – call it the file of interest (FOI) – that contains most of the data we wish to analyse, often derived from a survey. There is also a secondary file referred to as the linking file (LF) that contains additional variables on the individuals in the FOI. We have identification information on the individuals in both files and we use this to 'link' each individual in the FOI to the same individual in the LF. In some cases, hopefully the majority, the link is unequivocal and we can carry across the required variable values from the LF to the FOI. In other cases, however, there is some uncertainty about the link, due possibly to coding errors, naming ambiguities etc. It is in these cases that 'probabilistic data linkage models' are used and they estimate a set of weights w_{ij} for each individual, i, with an equivocal link. The subscript j indexes the records in the LF, where typically many of the w_{ij} will be zero indicating that these records are ruled out of consideration. For each i a lower cut-off is chosen for these weights to satisfy criteria related to specificity and specificity. If there remain w_{ij} above the cut-off the maximum of these is typically chosen and the corresponding records are regarded as linked.

One of the problems with such record linkage procedures is that when a probabilistic match is chosen there will remain some uncertainty about the correctness of the link and hence the values of the variables carried into the FOI. This 'identification error' should then be introduced into subsequent analyses, but this rarely occurs. Likewise, if no record in the LF has a weight in excess of the cut-off, some information about possible values of the variables will be lost. For both these reasons, traditional probabilistic methods can introduce bias into subsequent data modelling by underestimating the uncertainty associated with parameter estimates; this will be particularly important if many records are linked probabilistically.

We can use the idea of prior informed imputation to capture all the information produced by probabilistic record linkage.

16.8.2 Estimating a probability distribution using record linkage weights

For each record in the file of interest (FOI), each variable of interest (VOI) that has a missing value is considered in turn. For each such VOI we require a probability distribution (PD) for each possible value it can take, and this is derived from the linking file (LF). This distribution will then be used as a 'prior' distribution for the variable data value in the imputation, where we have been unable to establish the correct value through an unequivocal link.

In practice, we will generally only have the data available from the LF to estimate these prior distributions. We propose the following procedure. We assume that the weights are independent of the VOI. If not, our procedure will generally induce a (weak) dependency between the prior distributions and the distribution of the variable. At the imputation stage, this means that the distribution of the variable enters both

the prior and the standard imputation, which will generally lead to biased posterior selection probabilities.

From the probabilistic record linkage, for each record, *i,* in the FOI where we have not been able to find an unequivocal match, we have the set of weights $w_{ij}(j = 1, \ldots, m)$ over the m records indexed by j in the LF. In general many of these will be zero, for example, where a 'correct' match has already occurred.

Suppose first, that the variable of interest (VOI) z is discrete. For each possible value, k, of z we compute

$$p_{ik}(z) = c \sum_j w_{ij}\delta_{ik} / \sum_j w_{ij}$$
$$\delta_{ik} = 1 \quad if \quad z_i = k, 0 \text{ otherwise}$$

The scaling term c is chosen so that $\sum_k p_{ik}(z) = 1$ and this therefore provides a prior distribution for the discrete variable z.

For a continuous variable z we proceed as follows. For each record in the FOI we require the (weighted) distribution of z in the LF, and this distribution can vary from record to record. We therefore summarise each distribution by choosing a fixed set of intervals (s) with corresponding proportions for each FOI record. The number of intervals should be large enough to ensure a satisfactory number of intervals with nonzero probabilities, for each FOI record. Thus, for each record in the FOI we need to estimate a weighted frequency distribution for z in the LF and then determine $s-1$ cut points to produce the required set of intervals. A suitable weighted kernel density estimator could be used for this, or it could be approximated by defining a large number of intervals based upon the distribution of z in the LF and forming a weighted frequency count using the w_{ij}. This then provides a prior distribution for the continuous variable z.

Note that with multilevel data, where the variable is measured on a higher level unit, there is just one weight associated with each such higher level unit in the LF.

16.9 Conclusions

A very general method for dealing with fully or partially missing multilevel data has been described. The requirement to impute multiple complete datasets, often 20 or so are needed in practice, leads to significant amounts of computational time. While this is a drawback, it does mean that the datasets are available for repeated analyses with different models of interest that involve the variables used in the imputation.

We have discussed how MI can be used for data values that are 'partially known' and given a detailed discussion of an application in record linkage.

An alternative approach to multiple imputation using a joint model is the multiple imputation chained equation (MICE) approach, in which a set of conditional univariate models are used for imputation without specifying a joint model (Van Buuren, 2007). By contrast to the joint model approach it does not have a well established theoretical grounding, neither does it naturally extend to multilevel structures.

Finally, we could consider a full Bayesian approach to handling missing data where the imputation model is run in parallel with the MOI, so that at each MCMC cycle missing values are imputed and these imputed values are then used in the steps for the MOI. This method can be computationally very slow. Multiple imputation, an approximation to a full Bayesian model, uses frequentist properties for inference and has been implemented in software that can handle moderately large datasets.

17

Multilevel models with correlated random effects

17.1 Introduction

We have considered models where we assume that the residuals or random effects associated with an explanatory variable and belonging to the same classification, are independent. We now look at the effect of removing the independence assumption between these in the context of a 2-level model. We first consider removing the independence between the level 2 residuals and instead assume that the vector of all residuals at the cluster level follows a general multivariate normal distribution. We then show how to remove the assumption of independence between level 1 residuals within particular clusters.

17.2 Non-independence of level 2 residuals

Removing the assumption of independence between the level 2 random effects may be driven by the belief that some pairs of clusters are more similar to each other than to other clusters. Such a belief is what often drives spatial modelling, discussed in Chapter 13, where the relative locations of clusters of data are expected to influence the correlation between them. The relationship between cluster effects is expressed in terms of conditional rather than joint distributions as we do here but we note the similarities.

Another way in which non-independence occurs is when clusters themselves can be clustered into further higher level clusters; for example, in education pupils may be clustered into schools which themselves are clustered into education authorities. The education authority will often be fitted as an additional level of random effects

Multilevel Statistical Models: 4th Edition Harvey Goldstein

but alternatively can be accounted for in a 2-level modelling framework by removing some of the independence assumptions at level 2.

We begin by describing how an independence assumption at level 2 is removed. If we consider the standard variance components model

$$\begin{aligned} &y_{ij} = (X\beta)_{ij} + u_j + e_{ij} \\ &E(u_j) = 0, \text{var}(u_j) = \sigma_u^2, E(e_{ij}) = 0, \text{var}(e_{ij}) = \sigma_e^2 \\ &i = 1, \ldots, n_j, j = 1, \ldots, J \end{aligned} \tag{17.1}$$

and then let $\mathbf{u} = (u_1^T, u_2^T, \ldots., u_J^T)^T$, that is, all the residuals at level 2 stacked as a vector of length J. A more general model will then have $\mathbf{u} \sim MVN(0, \mathbf{\Omega}_u)$with the earlier independence assumption a special case with Ω_u diagonal, and with σ_u^2 on the diagonal. We now have flexibility in how we parameterise the covariance matrix Ω_u. We could consider an unconstrained representation and estimate all parameters in the covariance matrix but this will result in $J \times (J-1)/2$ parameters, although we shall see that this may be a model of interest in the case of level 1 residuals.

There may also be identifiability issues as, for example, if we assume that there is just a random intercept for each cluster and a different variance for each cluster, then in general there is insufficient data to identify these different variances without making some further assumptions about the variances for example by expressing an informative prior for them. We may also wish to model the variance as a function of further explanatory variables and we shall discuss this later.

For now, we will assume equal variances and write $\Omega_u = \sigma_u^2 D_u$ where D_u is the correlation matrix of the u's and σ_u^2 is the common variance term. We also write $\rho_{j_1 j_2}$ to represent the correlation between the effects for clusters j_1 and j_2. We now model the corresponding correlations using a functional form $f^{-1}(j_1, j_2, \boldsymbol{\alpha})$, which involves a set of distance measures for the level 2 units j_1 and j_2 and a set of parameters $\boldsymbol{\alpha}$. We shall assume here a generalised linear function of the form

$$f(\rho_{j_1 j_2}) = \alpha_1 g_1(j_1, j_2) + \alpha_2 g_2(j_1, j_2) + \ldots. \tag{17.2}$$

One choice of functional form is the inverse hyperbolic function

$$\begin{aligned} f_{j_1 j_2} &= f(\rho_{j_1 j_2}) = 2\tanh^{-1}(\rho_{j_1 j_2}) = \log\left(\frac{1+\rho_{j_1 j_2}}{1-\rho_{j_1 j_2}}\right) \\ \rho_{j_1 j_2} &= (e^{f_{j_1 j_2}} - 1)/(e^{f_{j_1 j_2}} + 1) \end{aligned}$$

which is effectively the Fisher z-transformation for a correlation coefficient and where the $g_h, h = 1 \ldots p$, are known. This function ensures that each correlation lies in the interval (–1, 1) (although of itself this does not guarantee that the final covariance matrix is positive definite):

Where we have independence between two random effects, these functions can be given values of zero. This can be achieved by introducing an indicator vector, $\delta_{j,k}$ to produce a final correlation structure defined by $\delta_{j,k}\rho_{jk}$.

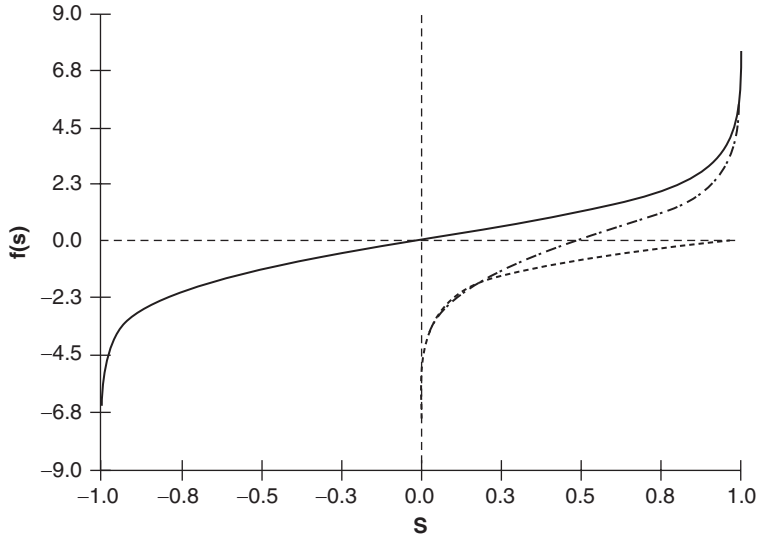


Figure 17.1 Link function f(s). From left to right: hyperbolic; logit; log.

An alternative link function, if we wish to restrict the correlations to be positive is the logit given by

$$\rho_{jk} = e^{f_{jk}}/(e^{f_{jk}} + 1)$$

or the log link function given by

$$\rho_{jk} = e^{f_{jk}}$$

In this last case, the correlations are positive and f is restricted to be negative to ensure the correlations are less than one. This constraint is achieved in the MCMC algorithm to be described, by only accepting negative values. We used the log link in Section 5.4 to describe an exponential decay function for repeated measures level 1 residuals, and discussed maximum likelihood estimates. The different functions are sketched in Figure 17.1.

17.3 MCMC estimation for non-independent level 2 residuals

We now consider how to sample the residuals and the level 2 covariance matrix. The steps for the fixed coefficients and level 1 variance are as in the standard case.

17.3.1 Sampling the level 2 covariance matrix

A uniform prior is assumed for the variance and for the $\{\alpha\}$.

We use $\Omega_{u[l,m]}$ to denote the *l,m*-th element of the level 2 covariance matrix with $\Omega_{u[l,l]} = \sigma_u^2$. We update the parameters $\{\alpha\}$, one at a time, using a Metropolis step and a normal random walk proposal as follows. We consider first the variances and then the $\{\alpha\}$.

At iteration t generate $\Omega^*_{u[l,l]} \sim \mathrm{N}(\Omega^{(t-1)}_{u[l,l]}, \sigma^2_{pl})$ where σ^2_{pl} is a proposal distribution variance that has to be set. Then form a proposed new matrix Ω^*_u by replacing the variance by this proposed value unless Ω^*_u is not positive definite in which case set $\Omega^{(t)}_{u[l,l]} = \Omega^{(t-1)}_{u[l,l]}$. We also update the covariances using the proposed value, using the current correlations determined by$\{\alpha\}$, with the new proposed variance. That is set $\Omega^t_{u[l,l]} = \Omega^*_{u[l,l]}$ with probability min[1, $p(\Omega^*_u|u)/p(\Omega^{(t-1)}_u|u)$] and $\Omega^{(t)}_{u[l,m]} = \Omega^{(t-1)}_{u[l,m]}$ otherwise.

The components of the likelihood ratio are

$$\begin{aligned} p(\Omega^*_u|u) &= |\Omega^*_u|^{-1/2} \exp -(u^T(\Omega^*_u)^{-1}u/2) \\ p(\Omega^{(t-1)}_u|u) &= |\Omega^{(t-1)}_u|^{-1/2} \exp -(u^T(\Omega^{(t-1)}_u)^{-1}u/2) \end{aligned} \tag{17.3}$$

with current estimates substituted.

For each element of $\{\alpha\}$ at iteration t generate $\alpha^*_l \sim \mathrm{N}(\alpha^{(t-1)}_l, \sigma^2_{\alpha,l})$ where $\sigma^2_{\alpha,l}$ is a proposal distribution variance that has to be set for each parameter. Then form a proposed new matrix Ω^*_u by replacing each element of $\Omega^{(t-1)}_u$ by the value computed from the updated set $\{\alpha\}$ using (2), unless Ω^*_u is not positive definite in which case set $\Omega^{(t)} = \Omega^{(t-1)}_u$. That is set $\Omega^{(t)}_u = \Omega^*_u$ with probability min[1, $p(\Omega^*_u|u)/p(\Omega^{(t-1)}_u|u)$] and $\Omega^{(t)}_u = \Omega^{(t-1)}_u$ otherwise.

The components of the likelihood ratio are, as in (17.3),

$$\begin{aligned} p(\Omega^*_u|u) &= |\Omega^*_u|^{-1/2} \exp -(u^T(\Omega^*_u)^{-1}u/2) \\ p(\Omega^{(t-1)}_u|u) &= |\Omega^{(t-1)}_u|^{-1/2} \exp -(u^T(\Omega^{(t-1)}_u)^{-1}u/2) \end{aligned}$$

The proposal distribution parameters can be chosen by an adaptive sampling procedure (see below).

To model additionally the variance as a function of explanatory variables, we write

$$\sigma^2_{uj} = \exp\left(\sum_{h=0}^{q} \gamma_h z_{hj}\right) \tag{17.4}$$

where the z_h are level 2 (or higher level) predictors and typically $z_0 = 1$ and the exponential form is chosen to ensure positive variances.

We therefore replace the step for updating the variance (17.3) with a series of MH steps, using suitable proposal distributions, for updating the γ_h. For each proposal we need, as before, to check that the resulting covariance matrix Ω_u^* is positive definite.

17.3.2 Sampling the level 2 residuals

The posterior distribution for the level 2 residuals is as follows

$$\begin{aligned}
&p(u|y, \Omega_u, \sigma_e^2) \propto \\
&\left(\tfrac{1}{\sigma_e^2}\right)^{N/2} \exp\left[-\frac{1}{2\sigma_e^2}(y-(X\beta)-(Zu))^T(y-(X\beta)-(Zu)) + \sigma_e^2 u^T \Omega_u^{-1} u\right] \\
&\text{so that we now sample from} \\
&u \sim N(\hat{u}, \hat{D})
\end{aligned} \tag{17.5}$$

where for the variance components case we have

$$\begin{aligned}
\hat{D} &= \sigma_e^2 \left[\sum_{i,j} Z_{ij}^T Z_{ij} + \sigma_e^2 \Omega_u^{-1}\right]^{-1} = \sigma_e^2 \left[diag(n_j) + \sigma_e^2 \Omega_u^{-1}\right]^{-1} \\
\hat{u} &= \left[\sum_{i,j} Z_{ij}^T Z_{ij} + \sigma_e^2 \Omega_u^{-1}\right]^{-1} \left[\sum_{i,j} Z_{ij}^T (y_{ij} - (X\beta)_{ij})\right] = \hat{D}\sigma_e^{-2}\tilde{y} \\
\tilde{y} &= \{\tilde{y}_j\}, \tilde{y}_j = \sum_i^{n_j} (y_{ij} - (X\beta)_{ij})
\end{aligned} \tag{17.6}$$

17.4 Adaptive proposal distributions in MCMC estimation

The proposal distributions are determined adaptively (Browne and Draper, 2006).

We choose a desirable acceptance rate r, say, 0.5, and we choose a batch size B, for example, 100, as used by Browne and Draper (2006). For each batch of iterations during the burn in period we compute the acceptance rate r^*. We update the proposal distribution, for each parameter, according to the following rule, from suitable starting values supplied by the user.

$$\text{If } r^* \geq r, \quad \theta_t = \theta_{t-1}[2 - \left(\frac{1-r^*}{1-r}\right)], \quad \text{otherwise} \quad \theta_t = \theta_{t-1}(2 - \frac{r^*}{r})^{-1} \tag{17.7}$$

where r is the desired acceptance rate and θ_t is the normal proposal distribution standard deviation for the parameter under consideration at iteration t. The proposal distribution adapting can be run for the duration of the burn in period.

17.5 MCMC estimation for non-independent level 1 residuals

Our general 2-level model now becomes

$$y_{ij} = (X\beta)_{ij} + (Zu)_j + e_{ij}, \quad e_j \sim MVN(0, \Omega_{ej}), \quad u \sim MVN(0, \Omega_u),$$
$$e_j = \{e_{ij}\} \tag{17.8}$$

where we assume that there is just a single residual term at level 1.

We assume the following form for the correlations, similar to those at level 2

$$f(\rho_{jk}^{(1)}) = \alpha_1^{(1)} g_1^{(1)}(t_j, t_k) + \alpha_2^{(1)} g_2^{(1)}(t_j, t_k) \ldots\ldots \text{and}$$
$$f(s) = 2\tanh^{-1}(s) = \log\left(\frac{1+s}{1-s}\right) \tag{17.9}$$

where the superscript (1) denotes level 1 and we can also have logit or log link functions.

The sampling for the level 1 covariance matrix involves essentially the same steps as for the level 2 matrix described above. For sampling the variance and correlation parameters the components of the likelihood ratio become

$$p(\Omega_e^*|e) = \prod_j |\Omega_{ej}^*|^{-1/2} \exp -(e_j^T(\Omega_{ej}^*)^{-1} e_j/2) \text{ and}$$
$$p(\Omega_e^{(t-1)}|e) = \prod_j |\Omega_{ej}^{(t-1)}|^{-1/2} \exp -(e_j^T(\Omega_{ej}^{(t-1)})^{-1} e_j/2)$$

The explanatory variables for the correlations $g_k^{(1)}(k = 1\ldots)$ must be specified for each level 2 unit.

When sampling the fixed effects, since the level 1 residuals are no longer independent, the standard MCMC step is modified as follows.

We assume a 'diffuse' prior $p(\beta) \propto 1$ so that

$$p(\beta|y, \Omega_e, u) \propto \prod_j |\Omega_{ej}|^{-1/2} \exp[-\tilde{y}_j^T \Omega_{ej}^{-1} \tilde{y}_j/2]$$
where $\tilde{y}_j = \{\tilde{y}_{ij}\}, \tilde{y}_{ij} = y_{ij} - (X\beta)_{ij} - (Zu)_{ij}$
so that we sample from
$$\beta \sim MVN(\hat{\beta}, \hat{D}_\beta) \tag{17.10}$$
$$\hat{D}_\beta = \left[\sum_j X_j^T \Omega_{ej}^{-1} X_j\right]^{-1}$$
$$\hat{\beta} = \hat{D}_\beta \left[X_j^T \Omega_{ej}^{-1}(y_j - (Zu)_j)\right]$$

Likewise, when sampling the level 2 random effects (17.7) becomes

$$\begin{aligned}\hat{D}_u &= \left[diag(Z_j^T \Omega_{ej}^{-1} Z_j) + \Omega_u^{-1}\right]^{-1} \\ \hat{u} &= \hat{D}_u \left[\{Z_j^T \Omega_{ej}^{-1}(y_j - (X\beta)_j)\}\right]\end{aligned} \tag{17.11}$$

17.5.1 A 2-level model formulated as a single level model with non-independent residuals

We now show how we may choose a particular non-independence pattern to specify a 2-level model using a single level formulation. Consider a 2-level variance components model. For a level 2 unit with four level 1 units, the covariance among these is given by

$$\begin{pmatrix} \sigma_e^2 + \sigma_u^2 & & & \\ \sigma_u^2 & \sigma_e^2 + \sigma_u^2 & & \\ \sigma_u^2 & \sigma_u^2 & \sigma_e^2 + \sigma_u^2 & \\ \sigma_u^2 & \sigma_u^2 & \sigma_u^2 & \sigma_e^2 + \sigma_u^2 \end{pmatrix} \tag{17.12}$$

For a single level model with an equal correlation structure we can write the covariance among the four level 1 units as

$$\begin{pmatrix} \sigma_{e*}^2 & & & \\ \rho\sigma_{e*}^2 & \sigma_{e*}^2 & & \\ \rho\sigma_{e*}^2 & \rho\sigma_{e*}^2 & \sigma_{e*}^2 & \\ \rho\sigma_{e*}^2 & \rho\sigma_{e*}^2 & \rho\sigma_{e*}^2 & \sigma_{e*}^2 \end{pmatrix} \tag{17.13}$$

which has an equivalent structure with $\rho\sigma_{e*}^2 = \sigma_u^2$, $\sigma_{e*}^2 = \sigma_u^2 + \sigma_e^2$, and ρ can be interpreted in the usual way as the intra-unit correlation. Note that if we compute the DIC statistics for these two different representations we will obtain different values since we are actually fitting different models. Model (17.12) includes random effect parameters whereas (17.13) does not (Browne and Goldstein (2010) give a numerical example).

17.6 Modelling the level 1 variance as a function of explanatory variables with random effects

To model the variance as a function of explanatory variables, we write

$$\sigma_{eij}^2 = \exp(\sum_{h=0}^{q} \gamma_h z_{hij}) \tag{17.14}$$

where the z_h may be level 1 or higher level predictors and typically $z_0 = 1$. We replace the step for updating the variance with a series of MH steps, using suitable proposal distributions, for updating the γ_h. For each proposal we need to check that each of the covariance matrices Ω^*_{ej} remains positive definite.

If we add a simple independent level 2 random effect to (17.14) we obtain

$$\sigma^2_{eij} = \exp(\sum_{h=0}^{q} \gamma_h z_{hij} + u_{1j}) \tag{17.15}$$

and we shall assume

$$u_{1j} \sim N(0, \sigma^2_{u1})$$

with density function at iteration t , $\phi_t(u_{1j})$.

For each level 2 unit at iteration t we propose a new value u^*_{1j} with proposal distribution $N(u^{(t-1)}_{1j}, \sigma^{*^2}_{u1})$, where $\sigma^{*^2}_{u1}$ can be chosen as the mean over the burn-in period. We accept the proposed value according to the Metropolis ratio

$$\frac{p(\Omega^*_{ej}|e)}{p(\Omega^{(t-1)}_{ej}|e)} \times \frac{\phi(u^*_{1j})}{\phi(u^{(t-1)}_{1j})}$$

To sample σ^2_{u1} we can use a Gibbs step, as described in Appendix 2.5.

We can elaborate (17.15) by adding further random coefficients: this leads to the updating of each one using an appropriate set of proposals. We can also allow the u_{1j} to be correlated with the other level 2 random effects. In this case, we can update the u_{1j} conditionally given the current values of the remaining level 2 random effects, in a similar fashion to the conditional steps described in Chapter 7. Hedeker *et al.* (2008) consider this model and show how to obtain maximum likelihood estimates.

A further possible elaboration is to allow the correlations to include random effects. Thus, we could rewrite (17.2) as

$$\begin{aligned} f(\rho^{(1)}_{j_1 j_2}) &= \alpha_1 g_1(j_1, j_2) + \alpha_2 g_2(j_1, j_2) + \ldots + u_{2j} \\ u_{2j} &\sim N(0, \sigma^2_{u2}) \end{aligned} \tag{17.16}$$

where the u_{1j} and u_{2j} may be correlated and also correlated with other level 2 random effects. Sampling is similar to that for the variance with a random effect.

17.7 Discrete responses with correlated random effects

These procedures can be extended to handle discrete responses. Consider the level 1 binary response case in detail, using a probit link function as described in Section 4.7.

For the case of non-independence at level 1, we modify the step to sample the latent normal variable as follows.

For level 2 unit *j,* suppose that the values at the start of the current iteration of the latent normal level 1 random effects (residuals), are given by $e_j = \{e_{ij}\}$. We sample each random effect conditioning on the remaining level 1 random effects in the level 2 unit. That is, for the conditional sampling, for level 1 unit k in level 2 unit j, the distribution $N[(X\beta)_{kj} + (ZU)_{kj}, 1]$ is replaced by

$$N[(X\beta)_{kj} + (ZU)_{kj} + e_{k,j}^T \Sigma_{12}^T \Sigma_1^{-1}, \sigma_k^2], \sigma_k^2 = 1 - \Sigma_{12}^T \Sigma_1^{-1} \Sigma_{12}, e_{k,j} = e_{j(j \neq k)}$$

where we write $\Omega_{ej} = \begin{pmatrix} \Sigma_1 & \\ \Sigma_{12} & 1 \end{pmatrix}$. This sequence will produce a multivariate normal distribution for the latent variables within each level 2 unit.

When sampling conditionally on correlated random effects, it is possible for the conditional mean to become relatively large and the corresponding residual variance to become relatively small so that for some data points we may be sampling from the extreme tail of the normal distribution. Given machine accuracy, this may lead to the associated tail probability being returned as 1.0 leading to a latent variable value that is coded as infinite. To avoid this problem, a cut-off should be chosen; for example, a value equivalent to 5 on the standard normal scale. Further details are given by Browne and Goldstein (2010).

17.7.1 The probit ordered response model where the variance is a function of explanatory variables with random effects

In Section 11.12.4, we described an ordered response probit model with threshold parameters (α_h) that were further modelled as functions of explanatory variables, namely

$$\begin{aligned} &\gamma^{(h)} = \int_{-\infty}^{\alpha_h - X\beta} \phi(z)dz, \quad \gamma^{(h)} = \sum_{g=1}^{h} \pi_g \\ &\pi_g = pr(t_{g-1} < t \leq t_g), \quad \alpha_h = \sum_{k=1}^{h} \exp(\alpha_k^* + \sum_{s=1}^{q} \delta_s z_{ks}), \quad \alpha_1^* = 0 \end{aligned} \tag{17.17}$$

where the exponential formulation for the threshold parameters guarantees monotonicity. Where the ordered response is at level 1 we can add one or more level 2 (or higher) random effects $\{u_j\}$ that are correlated with other level 2 random effects in the model. These may be sampled in the same way as with continuous responses.

Allowing the threshold parameters to depend on further variables implies that the latent normal scale itself will depend on such variables. This happens additively, with an increment added to each threshold value. As we have seen, it also allows for threshold specific explanatory variables as in the case of survival models. An alternative, multiplicative, scaling procedure is to specify the latent normal variance to be a function of explanatory variables, rather than being constrained to be 1. We

can write the variance in the form as in (17.15) as

$$\sigma^2_{eij} = \exp(\sum_{h=0}^{q} \gamma_h z_{hij} + u_{1j})$$

where we also allow one or more level 2 random effects, u_{1j}. Such a model has been used in the context of ROC analysis that studies the relationship between the false positive and true positive rates for a diagnostic test. See, for example, Ishwaran and Gatsonis (2000) who describe such a model but without random effects. Hedeker *et al*. (2009) describe a model that includes random effects, but with a logistic link function.

We may also consider models where we model both the thresholds and the variance, using different explanatory variables.

17.8 Calculating the DIC statistic

For a normal response, the likelihood for the j-th level 2 unit is given by

$$L_j = (2\pi)^{-n_j/2}|\Omega_{ej}|^{-1/2}\exp[-\tilde{y}_j^T\Omega_{ej}^{-1}\tilde{y}_j/2]$$
$$\text{where } \tilde{y}_j = \{\tilde{y}_{ij}\}, \quad \tilde{y}_{ij} = y_{ij} - (X\beta)_{ij} - (Zu)_{ij}$$

so that the total deviance is

$$-2Log(L) = N\log(2\pi) + \sum_j (\log|\Omega_{ej}| + \tilde{y}_j^T\Omega_{ej}^{-1}\tilde{y}_j)$$

In the case where the level 1 residuals are independent this becomes

$$-2Log(L) = N\log(2\pi) + N\log(\sigma_e^2) + \sum_{ij}\tilde{y}_{ij}^2/\sigma_e^2$$

The DIC statistic and the equivalent degrees of freedom are then computed in the usual way for each sample set of parameter values and for the mean parameter values as outlined in Section 2.13.4.

For a binary response the calculation is more complicated. Using the probit formulation, in the case of independent level 1 residuals we have, for the response y, given the parameter values

$$-2\log(L) = -2\sum_{ij}\log(p_{ij})$$
$$p_{ij} = \begin{cases} \int_{-\infty}^{(X\beta+Zu)_{ij}} \phi(t)dt, & y_{ij} = 1 \\ \int_{-\infty}^{-(X\beta+Zu)_{ij}} \phi(t)dt, & y_{ij} = 0 \end{cases}$$

where the level 1 residuals are dependent, using the joint likelihood for a level 2 unit involves the computation of a multidimensional integral, which could be carried out using simulation from the currently estimated multivariate normal distribution.

17.9 A growth dataset

We reanalyse the repeated measures data of Chapter 5. There, a 2-level polynomial growth model was fitted with terms up to the fourth order and with level 2 random coefficients for the intercept, linear and quadratic coefficients, and we fit the same model here. At level 1, the following link function is used to describe the correlation structure.

$$f_{t_1t_2} = \alpha|t_1 - t_2|, \quad \rho_{t_1t_2} = \exp(-f_{t_1t_2})$$

Table 17.1 shows the results from fitting the model using the inverse tanh link.

The estimate of alpha is close to zero, and for the log and logit links it is a large positive value. This contrasts with the maximum likelihood estimate of 6.9 with SE of 2.0. Figure 17.2 shows the chain for alpha.

Table 17.1 Height as a fourth degree polynomial on age, measured about 13.0 years. Standard errors in brackets. MCMC estimates. Burn-in = 5000, Sample = 50 000. After adapting, proposal distribution SD = 0.08.

Fixed			
Intercept	148.9 (1.3)		
age	6.16 (0.35)		
age^2	2.16 (0.47)		
age^3	0.39 (0.16)		
age^4	−1.55 (0.46)		
cos (time)	−0.24 (0.07)		
Random			
level 2 covariance matrix			
	Intercept	*Age*	*Age squared*
Intercept	65.9 (19.7)		
age	8.5 (3.5)	3.0 (0.9)	
Age squared	1.5 (1.6)	0.9 (0.4)	0.64 (0.25)
level 1 variance			
σ_e^2	0.21 (0.03)		
α (mean)	0.020 (0.20)		
α (median + 95 % interval)	0.064 (–0.42, 0.28)		
DIC (PD)	344.0 (58.1)		

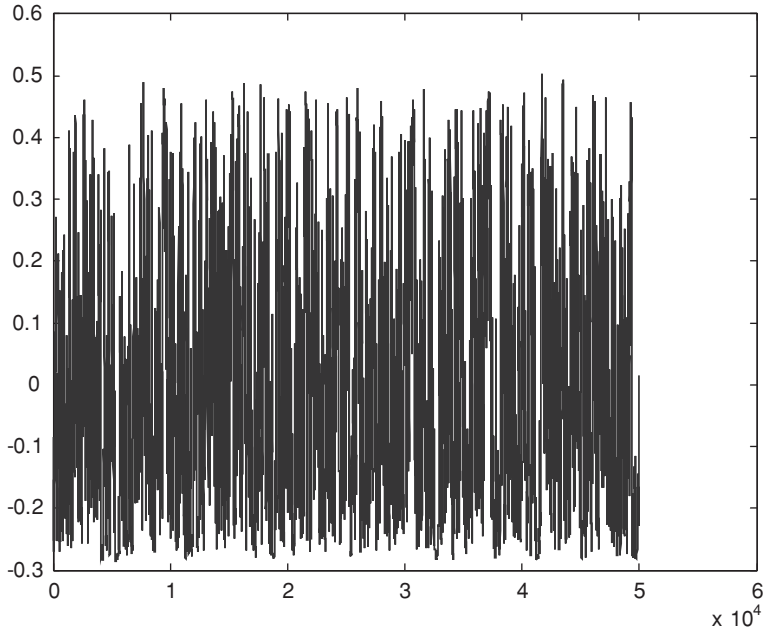


Figure 17.2 Chain for alpha in Table 17.1.

17.10 Conclusions

There are several extensions to these models, including higher levels of the data hierarchy and cross classifications where independence across classifications is assumed and also to multiple membership models. The above estimation steps would be applied to each relevant classification conditional on current estimates. In addition, we can consider a fully multivariate version with a set joint responses. The choice of prior distribution for the elements of the covariance matrix could be changed so that we could choose an inverse Gamma prior for the variance parameters, or a more informative prior.

A number of models can be viewed as special cases. Time series models such as autoregressive structures (Goldstein *et al*., 1994) are one example and within-family sibling relationship models are another. In the latter case, the correlation between sibling characteristics will typically depend on whether they are twins or singletons, or on whether they are half or full siblings. In the former case, an important feature of our model is that we can model complex time-series structures where the repeated measures do not occur at the same set of regular time intervals for each individual. This distinguishes it from standard SEM approaches that treat the set of common occasions as a special kind of multivariate structure.

Another important application of these models is in the modelling of educational and other data where institutions do not behave independently. Thus, for example, in the case of schooling effects, actions taken by one school when competing for limited resources can be expected to impact on the actions of nearby schools, and partnerships and collaborations will also invalidate assumptions about the independence of school effects on pupil progress or performance.

18

Software for multilevel modelling

18.1 Software packages

There are now many software packages that will carry out multilevel modelling. Most of the major statistical packages have features for the basic models, and many can fit more complex or specialised models. We shall not give a detailed review here, but merely list the main packages together with web addresses for further information (all were accessed in June 2010). In addition, there is much software written in general purpose programming languages such as C++, MATLAB, R etc. Published reviews of some of the packages are those of De Leeuw and Kreft (2001), Zhou *et al*. (1999) and Fein and Lissitz (2000).

The Centre for Multilevel Modelling (http://www.cmm.bristol.ac.uk/index.shtml) has a series of reviews, links and details of training sessions and workshops. It is also developing a set of web based training materials in multilevel modelling. Another good general site for resources is http://statcomp.ats.ucla.edu/mlm/.

There is a very active email discussion group that can be accessed and joined at www.jiscmail.ac.uk/lists/multilevel.html. The group serves as a means of exchanging information and suggestions about data analysis. Table 18.1 lists packages, together with the internet address and a brief note.

Multilevel Statistical Models: 4th Edition Harvey Goldstein

Table 18.1 Some available software packages.

Name	Web site	Note
aML	www.applied-ml.com	Concentrates on event history and multiprocess models. Maximum likelihood estimation.
ASREML	http://www.vsni.co.uk/software/asreml/	Same features as GENSTAT (below).
BAYESX	www.stat.uni-meunchen.de/~lang/bayesx/bayesx.html	General purpose MCMC estimation. Continuous and discrete responses with nested and cross classified structures. Concentrates on semiparametric regression.
BMDP	http://www.statistical-solutions-software.com/products-page/bmdp-statistical-software/	Variance components model and serial correlations for nested structures. Maximum likelihood and GEE.
EGRET	www.cytel.com/products/egret	Discrete responses for nested structures up to two levels. Maximum likelihood estimation.
GENSTAT	www.nag.co.uk/stats/tt_soft.asp	Continuous and discrete responses. Nested and cross-classified structures. Maximum likelihood estimation.
GLLAMM	www.gllamm.org/	Wide class of models including multilevel structural equation models with discrete and continuous responses. Maximum likelihood estimation. See also STATA.
HLM	http://www.ssicentral.com/hlm/	Continuous and discrete responses up to three levels. Serial correlation structures; measurement errors. Maximum and quasilikelihood estimation.
Latent GOLD	http://www.statisticalinnovations.com/products/latentgold_v4.html	Continuous and discrete responses. Multilevel structural equations. Mixture models. Up to three levels. Maximum likelihood estimation.
LIMDEP	www.limdep.com	General purpose econometric software. Continuous and discrete responses. Nested structures. Maximum likelihood estimation.
LISREL	http://www.ssicentral.com/lisrel/index.htm	Multilevel structural equations. Nested data structures. Maximum likelihood estimation.

MIXOR, MIXREG	http://tigger.uic.edu/~hedeker/mix.html	A suite of programs for continuous and discrete response multicategory models up to three levels. Maximum likelihood estimation.
MLwiN	http://www.cmm.bristol.ac.uk/index.shtml	General purpose package. Continuous and discrete responses for nested, cross classified and multiple membership structures for any number of levels. Serial correlation structures; event history models; factor analysis; measurement errors. Maximum and quasi likelihood estimation; MCMC estimation.
MPLUS	www.statmodel.com/	Continuous and discrete responses. Nested structures. Multilevel structural equations. Maximum likelihood estimation.
OSWALD	http://www.maths.lancs.ac.uk/Software/Oswald/	Works with S Plus for analysis of serial correlation and event history data. Maximum likelihood and GEE estimation.
REALCOM REALCOM-IMPUTE	http://www.cmm.bristol.ac.uk/realcom	Two level multivariate responses where these can be normal or discrete, using a latent normal model. REALCOM-IMPUTE interfaces with MLwiN for multiple imputation. MCMC estimation.
SAS	http://www.sas.com/technologies/analytics/statistics/stat/index.html	General purpose package. Continuous and discrete responses for nested and cross-classified structures Serial correlation structures. Maximum and quasi likelihood estimation for generalised linear models. MCMC estimation.
S-PLUS 2000	www.insightful.com	General purpose software. Continuous and discrete responses for nested structures.
SPSS (PASW package)	www.spss.com/	General purpose package, now owned by IBM. Handles basic continuous and discrete response models for nested structures. Autoregressive level 1 models. Maximum likelihood estimation
STATA	www.stata.com	General purpose package. Continuous and discrete responses. Nested and cross classified structures up to two levels. Structural equation models (GLLAMM program), serial correlations. Maximum likelihood estimation.
SYSTAT	http://www.spssscience.com/systat	General purpose package. Continuous responses. Nested structures up to two levels. Serial correlations. Maximum likelihood estimation.
WINBUGS	www.mrc-bsu.cam.ac.uk	General purpose package. Uses MCMC to fit a very wide range of models via a statistical control language. Continuous and discrete responses, measurement errors, factor analysis, serial correlations and more.

References

Abdous, B. and Berlinet, A. (1998) Pointwise improvement of multivariate kernel density estimates. *Journal of Multivariate Analysis*, **65**, 109–128.

Afshartous, D. and Wolf, M. (2007) Avoiding 'data snooping' in multilevel and mixed effects models. *Journal of the Royal Statistical Society*, *Series A*, **170**, 1035–1060.

Aitchison, J. and Bennett, J.A. (1970) Polychotomous quantal response by maximum indicant. *Biometrika*, **57**, 253–262.

Aitkin, M. (2010) *Statistical Inference: An Integrated Bayesian/Likelihood Approach*, Chapman & Hall/CRC, London.

Aitkin, M. and Longford, N. (1986) Statistical modelling in school effectiveness studies (with discussion). *Journal of the Royal Statistical Society, Series A*, **149**, 1–43.

Aitkin, M., Anderson, D. and Hinde, J. (1981) Statistical modelling of data on teaching styles (with discussion). *Journal of the Royal Statistical Society, Series A*, **144**, 148–161.

Aitkin, M., Anderson, D., Francis, B. *et al.* (1989) *Statistical Modelling in GLIM*, Clarendon Press, Oxford.

Allen, R. and Vignoles, A. (2007) What should an index of school segregation measure? *Oxford Review of Education*, **33**, 643–668.

Ansari, A. and Jedidi, K. (2002) Heterogeneous factor analysis models: a Bayesian approach. *Psychometrika*, **67**, 49–78.

Barbosa, M.F. and Goldstein, H. (2000) Discrete response multilevel models for repeated measures: an application to voting intentions data. *Quality and Quantity*, **34**, 323–330.

Beale, E.M.L. and Little, R.J.A. (1975) Missing values in multivariate analysis. *Journal of the Royal Statistical Society, Series B*, **37**, 129–145.

Bennett, N. (1976) *Teaching Styles and Pupil Progress*, Open Books, London.

Blatchford, P., Goldstein, H. and Mortimore, P. (1998) Research on class size effects: a critique of methods and a way forward. *International Journal of Educational Research*, **29**, 691–710.

Blatchford, P., Goldstein, H., Martin, C. *et al.* (2002) A study of class size effects in English school reception year classes. *British Educational Research Journal*, **28**, 169–185.

Blossfeld, H.P., Golsch, K. and Rohwer, G. (2007) *Event History Analysis with Stata*, Lawrence Erlbaum Associates, New Jersey.

Bock, R.D. (1992) Structural and nonstructural analysis of multiphasic growth, University of Chicago, (unpublished).

Bollen, K.A. and Curran, P.J. (2006) *Latent Curve Models: A Structural Equation Approach*, John Wiley & Sons, Hoboken.

Box, G.E.P. and Cox, D.R. (1964) An analysis of transformations (with discussion). *Journal of the Royal Statistical Society, B Series*, **26**, 211–252.

Breslow, N.E. and Clayton, D.G. (1993) Approximate inference in generalised linear mixed models. *Journal of the American Statistical Association*, **88**, 9–25.

Browne, W.J. (1998) Applying MCMC methods to multilevel models. PhD Thesis, University of Bath.

Browne, W.J. and Draper, D. (2000) Implementation and performance issues in the Bayesian and likelihood fitting of multilevel models. *Computational statistics*, **15**, 391–420.

Browne, W.J. and Goldstein, H. (2010) MCMC sampling for a multilevel model with non-independent residuals within and between cluster units. *Journal of Educational and Behavioural Statistics*, (to appear).

Browne, W.J. (2008) *MCMC Estimation in MLwiN*, Institute of Education, London.

Browne, W.J. and Draper, D. (2006) A comparison of Bayesian and likelihood based methods for fitting multilevel models. *Bayesian Analysis*, **1**, 473–514.

Browne, W.J., Golalizadeh Lahi, M. and Parker, R.M.A. (2009) *A Guide to Sample Size Calculations for Random Effect Models via Simulation and the MLPowSim Software Package*, University of Bristol, Bristol.

Browne, W.J., Draper, D., Goldstein, H. *et al.* (2002) Bayesian and likelihood methods for fitting multilevel models with complex level 1 variation. *Computational Statistics and Data Analysis*, **39**, 203–225.

Browne, W.J, Goldstein, H. and Rasbash, J. (2001a) Multiple membership multiple classification (MMMC) models. *Statistical Modelling*, **1**, 103–124.

Browne, W., Goldstein, H., Woodhouse, G. *et al.* (2001b) An MCMC algorithm for adjusting for errors in variables in random slopes multilevel models. *Multilevel Modelling Newsletter*, **13** (1), 4–9.

Browne, W., Steele, F., Golalizadeh, M. *et al.* (2009) The use of simple reparameterisations to improve the efficiency of Markov Chain Monte Carlo estimation for multilevel models with applications to discrete time survival models. *Journal of the Royal Statistical Society, Series A*, **172** (3), 579–598.

Bryk, A.S. and Raudenbush, S.W. (2002) *Hierarchical Linear Models*, Sage, California.

Bryk, A.S., Raudenbush, S.W., Seltzer, M. *et al.* (1988) *An Introduction to HLM: Computer Program and User's Guide*, 2nd edn, University of Chicago Department of Education, Chicago.

Burstein, L., Fischer, K.H. and Miller, M.D. (1980) The multilevel effects of background on science achievement: a cross national comparison. *Sociology of Education*, **53**, 215–225.

Bynner, J., Elias, P., McKnight, A. *et al.* (2002) *Young People in Transition: Changing Pathways to Employment and Independence*, Joseph Rowntree Foundation, York.

Carpenter, J.R., Ho, T., Pocock, S. *et al.* (2003) Modelling a repeated ordered categorical response in clinical trials with smoothing splines: angina grade in the RITA–2 trial. Technical report. Medical Statistics Unit, London School of Hygiene and Tropical Medicine.

Carpenter, J., Goldstein, H. and Rasbash, J. (1999) A nonparametric bootstrap for multilevel models. *Multilevel modelling newsletter*, **11**, 2–5.

Carpenter, J., Goldstein, H. and Rasbash, J. (2003) A novel bootstrap procedure for assessing the relationship between class size and achievement. *Journal of the Royal Statistical Society, Series C*, **52**, 431–443.

Carpenter, J.R., Kenward, M.G. *et al.* (2006) A comparison of multiple imputation and doubly robust estimation for analyses with missing data. *Journal of the Royal Statistical Society, Series A*, **169**, 571–584.

Clayton, D. and Rasbash, J. (1999) Estimation in large crossed random effect models by data augmentation. *Journal of the Royal Statistical Society, Series A*, **162**, 425–436.

Clayton, D.G. (1988) The analysis of event history data: a review of progress and outstanding problems. *Statistics in Medicine*, **7**, 819–841.

Clayton, D.G. (1991) A Monte Carlo method for Bayesian inference in frailty models. *Biometrics*, **47**, 467–485.

Clayton, D.G. (1992) Bayesian analysis of frailty models, Cambridge, MRC Biostatistics Unit, (unpublished).

Clayton, D. and Kaldor, J. (1987) Empirical Bayes estimates of age-standardised relative risks for use in disease mapping. *Biometrics*, **43**, 671–81.

Cleveland, W.S. and Devlin, S.J. (1988) Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American Statistical Association*, **83**, 596–610.

Cochran, W.G. (1983) *Planning and Analysis of Observational Studies*, John Wiley & Sons, Inc., New York.

Cohen, M.P. (1998) Determining sample sizes for surveys with data analysed by hierarchical linear models. *Journal of Officia Statistics*, **14**, 267–275.

Cole, T.J. and Green, P.J. (1992) Smoothing reference centile curves: the LMS method and penalized likelihood. *Statistics in Medicine*, **11**, 1305–1319.

Cook, R.D. and Weisberg, S. (1982) *Residuals and Influenc in Regression*, Chapman and Hall, London.

Courgeau, D., ed. (2007) *Multilevel Synthesis*, Springer-Verlag, Dordrecht.

Cox, D.R. (1972) Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B*, **34**, 187–220.

Cox, D.R. and Oakes, D. (1984) *Analysis of Survival Data*, Chapman and Hall, London.

Craig, P. (2007) A new reconstruction of multivariate normal orthant probabilities. *Journal of the Royal Statistical Society, Series B*, **70**, 227–243.

Creswell, M. (1991) A multilevel Bivariate Model, in *Data Analysis with ML3*, (eds R. Prosser, J. Rasbash, and H. Goldstein), Institute of Education, London.

Cronbach, L.J. and Webb, N. (1975) Between class and within class effects in a repeated aptitude × treatment interaction: reanalysis of a study by G.L. Anderson. *Journal of Educational Psychology*, **67**, 717–724.

Crouchley, R. and Ganjali, M. (2002) The common structure of several models for non-ignorable dropout. *Statistical Modelling*, **2**, 39–62.

Davison, A.C. and Hinkley, D.V. (1997) *Bootstrap Methods and their Application*, Cambridge University Press, Cambridge.

De Leeuw, J. and Kreft, I. (2001) Software for multilevel analysis, in *Multilevel modelling of health statistics*, (eds A. Leyland and H. Goldstein), John Wiley & Sons Ltd, Chichester.

Demirjian, A., La Palme, L. and Thibault, H.W. (1982) La croissance staturo-ponderale des enfants Canadien-Francais de la naissance à 36 mois. *Union Med Can*, **112**, 153–163.

Derbyshire, M.E. (1987) Statistical rationale for grant-related expenditure assessment (GREA) concerning personal social services. *Journal of the Royal Statistical Society, Series A*, **150**, 309–333.

Diggle, P.J. (1988) An approach to the analysis of repeated measurements. *Biometrics*, **44**, 959–971.

Diggle, P.J. and Kenward, M.G. (1994) Informative dropout in longitudinal data analysis. *Applied Statistics*, **43**, 49–93.

Draper, D. (2002) *Bayesian Hierarchical Modeling*, Springer-Verlag, New York.

Ecob, R. and Goldstein, H. (1983) Instrumental variable methods for the estimation of test score reliability. *Journal of Educational Statistics*, **8**, 223–241.

Efron, B. (1988) Logistic regression, survival analysis, and the Kaplan-Meier curve. *Journal of the American Statistical Association*, **83**, 414–425.

Efron, B. and Gong, G. (1983) A leisurely look at the Bootstrap, the Jacknife and Cross-validation. *The American Statistician*, **37**, 36–48.

Egger, M., Davey Smith, G., Schneider, M. *et al.* (1997) Bias in meta analysis detected by a simple graphical test. *British Medical Journal*, **315**, 629–634.

Egger, P.J. (1992) Event history analysis: discrete-time models including unobserved heterogeneity, with applications to birth history data. University of Southampton, PhD thesis.

Everitt, B.S. (1984). *Introduction to Latent Variable Models*, Chapman & Hall, London.

Fein, M. and Lissitz, R.W. (2000) Comparison of HLM and MLwiN multilevel analysis software packages: a Monte Carlo investigation into the equality of the estimates. Working paper, University of Maryland.

Finner, H. and Gontscharuk, V. (2009) Controlling the familywise error rate with plug-in estimator for the proportion of true null hypotheses. *Journal of the Royal Statistical Society, Series B*, **71**, 1031–1048.

Friedman, J.H. and Silverman, B.W. (1989) Flexible parsimonious smoothing and additive modeling (with discussion). *Technometrics*, **31**, 3–21.

Fuller, W.A. (2006) *Measurement Error Models*, John Wiley & Sons Inc., New York.

Garrett, M., Fitzmaurice, M. and Laird, N. (1993) A likelihood based method for analysing longitudinal binary responses. *Biometrika*, **80**, 141–51.

Gelfand, A.E., Sahu, S.K. and Carlin, B.P. (1995) Efficient parameterisations for normal linear mixed models. *Biometrika*, **82**, 479–488.

Gelman, A. and Hill, J. (2007) *Data Analysis Using Regression and Multilevel/Hierarchical Models*, Cambridge University Press, New York.

Gelman, A., Roberts, G.O. and Gilks, W.R. (1996) Efficient Metropolis jumping rules. *Bayesian Statistics*, **5**, 599–607.

Gilks, W.R. and Wild, P. (1992) Adaptive rejection sampling for Gibbs sampling. *Applied Statistics*, **41**, 337–348.

Gilks, W.R., Richardson, S. and Spiegelhalter, D.J. (1996) *Markov Chain Monte Carlo in practice*. Chapman and Hall, London.

Gilks, W.R., Clayton, D.G., Spiegelhalter, D.J. *et al.* (1993) Modelling complexity: applications of Gibbs Sampling in medicine. (With discussion). *Journal of the Royal Statistical Society, Series B*, **55**, 39–102.

Goldstein, H. (1976) Smoking in pregnancy: some notes on the statistical controversy. *British Journal of Preventive and Social Medicine*, **31**, 13–17.

Goldstein, H. (1979) *The Design and Analysis of Longitudinal Studies*, Academic Press, London.

Goldstein, H. (1986) Multilevel mixed linear model analysis using iterative generalised least squares. *Biometrika*, **73**, 43–56.

Goldstein, H. (1987a) Multilevel covariance component models. *Biometrika*, **74**, 430–431.

Goldstein, H. (1987b) *Multilevel Models in Educational and Social Research*, Griffin, London.

Goldstein, H. (1987c) The choice of constraints in correspondence analysis. *Psychometrika*, **52**, 207–215.

Goldstein, H. (1989a) Restricted unbiased iterative generalised least squares estimation. *Biometrika*, **76**, 622–623.

Goldstein, H. (1989b) Efficient prediction models for adult height, in Auxology 88; *Perspectives in the Science of Growth and Development*, (ed J.M. Tanner), Smith Gordon, London.

Goldstein, H. (1991) Nonlinear multilevel models with an application to discrete response data. *Biometrika*, **78**, 45–51.

Goldstein, H. (1992) Statistical information and the measurement of education outcomes (editorial). *Journal of the Royal Statistical Society, Series A*, **155**, 313–315.

Goldstein, H. (1995) *Multilevel Statistical Models*, 2nd edn, Edward Arnold, London.

Goldstein, H. (1997) Methods in school effectiveness research. *School effectiveness and school improvement*, **8**, 369–395.

Goldstein, H. (1998) *Models for reality*, Institute of Education, London.

Goldstein, H. (2003) *Multilevel Statistical Models*, 3rd edn, Edward Arnold, London.

Goldstein, H. (2009) Handling attrition and non-response in longitudinal data. *International Journal of Longitudinal and Life Course Studies*, **1**, 63–72.

Goldstein, H. (2010a) A general model for the analysis of multilevel discrete time survival data. Submitted for publication.

Goldstein, H. (2010b) Estimating research performance using research grant award gradings. *Journal of the Royal Statistical Society, Series A*, (to appear).

Goldstein, H. and Blatchford, P. (1998) Class size and educational achievement: a review of methodology with particular reference to study design. *British Educational Research Journal* **24**, 255–268.

Goldstein, H. and Browne, W. (2002) Binary response factor modelling of achievement data, in *Psychometrics: A Festschrift to Roderick P. McDonald* (eds A. Olivares and J. McArdle), Lawrence Earlbaum, Hillsdale, N.J.

Goldstein, H. and Browne, W. (2005) Multilevel factor analysis models for continuous and discrete data, in *Contemporary Psychometricss*, (eds A. Mayden-Olivares and J.J. McArdle), Lawrence Erlbaum Associates, New Jersey.

Goldstein, H. and Noden, P. (2003) Modelling social segregation. *Oxford Review of Education*, **29**, 225–237.

Goldstein, H. and Rasbash, J. (1996) Improved approximations for multilevel models with binary responses. *Journal of the Royal Statistical Society, Series A*, **159**, 505–513.

Goldstein, H. and Spiegelhalter, D.J. (1996) League tables and their limitations: statistical issues in comparisons of institutional performance. *Journal of the Royal Statistical Society, Series A*, **159**, 385–443.

Goldstein, H. and Wood, R. (1989) Five decades of item response modelling. *British Journal of Mathematical and Statistical Psychology*, **42**, 139–167.

Goldstein, H., Bonnet, G. and Rocher, T. (2007a) Multilevel structural equation models for the analysis of comparative data on educational performance. *Journal of Educational and Behavioural Statistics*, **32**, 252–286.

Goldstein, H., Browne, W. and Rasbash, J. (2002) Partitioning variation in multilevel models. *Understanding Statistics*, **1**, 223–232.

Goldstein, H., Carpenter, J. and Gayle, V. (2010) Modelling missing data for longitudinal surveys: a weighted multiple imputation procedure. Submitted for publication.

Goldstein, H., Carpenter, J., Kenward, M. *et al.* (2009) Multilevel models with multivariate mixed response types. *Statistical Modelling*, **9** (3), 173–197.

Goldstein, H., Pan, H. and Bynner, J. (2004) A flexible procedure for analyzing longitudinal event histories using a multilevel model. *Understanding Statistics*, **3**, 85–89.

Goldstein, H., Rasbash, J., Browne, W. *et al.* (2000a) Multilevel models in the study of dynamic household structures. *European Journal of Population*, **16**, 373–387.

Goldstein, H., S. Burgess and B. McConnell (2007b) Modelling the impact of pupil mobility on school differences in educational achievement. *Journal of Royal Statistical Society, Series A*, **170**, 941–954.

Goldstein, H., Yang, M., Omar, R. *et al.* (2000b) Meta analysis using multilevel models with an application to the study of class size effects. *Journal of the Royal Statistical Society, Series C*, **49**, 399–412.

Goldstein, H. and McDonald, R.P. (1987) A general model for the analysis of multilevel data. *Psychometrika*, **53**, 455–467.

Goldstein, H. and Rasbash, J. (1992) Efficient computational procedures for the estimation of parameters in multilevel models based on iterative generalised least squares. *Computational Statistics and Data Analysis*, **13**, 63–71.

Goldstein, H. and Healy, M.J.R. (1995) The graphical presentation of a collection of means. *Journal of the Royal Statistical Society*, *Series A*, **158**, 175–177.

Goldstein, H., Healy, M.J.R. and Rasbash, J. (1994) Multilevel time series models with applications to repeated measures data. *Statistics in Medicine*, **13**, 1643–1655.

Goldstein, H., Kounali, D. and Robinson, A. (2008). Modelling measurement errors and category missclassifications in multilevel models. *Statistical Modelling* **8** (3), 243–261.

Goldstein, H., Rasbash, J., Yang, M. *et al.* (1993) A multilevel analysis of school examination results. *Oxford Review of Education*, **19**, 425–433.

Green, P.J. and Silverman, B.W. (1994) *Nonparametric Regression and Generalised Linear Models*, Chapman & Hall, London.

Greenacre, M.J. (1984) *Theory and Applications of Correspondence Analysis*, Academic Press, New York.

Grizzle, J.C. and Allen, D.M. (1969) An analysis of growth and dose response curves. *Biometrics*, **25**, 357–361.

Gumpertz, M.L. and Pantula, S.G. (1992) Nonlinear regression with variance components. *Journal of the American Statistical Association*, **87**, 201–209.

Harrison, G.A. and Brush, G. (1990) On correlations between adjacent velocities and accelerations in longitudinal growth data. *Annals of Human Biology*, **17**, 55–57.

Hartzel, J., Agresti, A. and Caffo, B. (2001) Multinomial logit random effects models. *Statistical Modelling* **1**, 81–102.

Hastie, T.J. and Tibshirani, R.J. (1990) *Generalized Additive Models*, Chapman & Hall, London.

Hawkes, D. and Plewis, I. (2006) Modelling non-response in the National Child Development Study. *Journal of the Royal Statistical Society, Series A*, **169**, 479–492.

Heagerty, P.J. and Zeger, S.L. (2000) Marginalized multilevel models and likelihood inference (with discussion). *Statistical Science*, **15**, 1–26.

Healy, M.J.R., Rasbash, J. and Yang, M. (1988) Distribution-free estimation of age-related centiles. *Annals of Human Biology*, **15**, 17–22.

Heath, A., Jowell, R., Curtice, J. *et al.* (1991) *Understanding Political Change*, Pergamon, Oxford.

Heath, A., Yang, M. and Goldstein, H. (1996) Multilevel analysis of the changing relationship between class and party in Britain 1964–1992. *Quality and Quantity*, **30**, 389–404.

Heck, R.H. and Thomas, S.L. (2000) *An introduction to multilevel modelling techniques*, Lawrence Erlbaum Associates, New Jersey.

Heckman, J.J. (1979) Sample selection bias as a specification error. *Econometrica*, **47**, 153–161.

Hedeker, D. and Gibbons, R.D. (1994) A random effects ordinal regression model for multilevel analysis. *Biometrics*, **50**, 933–944.

Hedeker, D., Hakan, D. and Mermelstein, J. (2009) A mixed ordinal scale model for analysis of ecological momentary assessment (EMA) data. *Statistics and its Interface*, **2**, 391–401.

Hedeker, D., Mermelstein, R.J. and Demirtas, H. (2008) An application of a mixed-effects location scale model for analysis of ecological momentary assessment (EMA) data. *Biometrics*, **64**, 627–634.

Hedges, L. (2009) Adjusting a significance test for clustering in designs with two levels of nesting. *Journal of Educational and Behavioral Statistics*, **34**, 1–27.

Hedges, L.V. and Olkin, I.O. (1985) *Statistical Methods for Meta Analysis*, Academic Press, Florida.

Heitjan, D.F. and Rubin, D.B. (1991) Ignorability and coarse data. *Annals of Statistics*, **19**, 2244–2253.

Higgins, J.P.T., Thompson, S.G. and Spiegelhalter, D.J. (2009) A re-evaluation of random effects meta analysis. *Journal of the Royal Statistical Society, Series A*, **172** (1), 137–159.

Hill, P.W. and Goldstein, H. (1998) Multilevel modelling of educational data with cross classification and missing identification of units. *Journal of Educational and Behavioural statistics*, **23**, 117–128.

Hodges, J. (1998) Some algebra and geometry for hierarchical models, applied to diagnostics. *Journal of the Royal Statistical Society, Series B*, **60**, 497–536.

Holland, P.W. (1986) Statistics and causal inference. *Journal of the American Statistical Association*, **81**, 945–971.

Holm, S. (1979) A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, **6**, 65–70.

Hox, J. (2002) *Multilevel Analysis, Techniques and Applications*, Lawrence Erlbaum Associates, New Jersey.

Hu, Z., Wang, N. and Carroll, R.J. (2004) Profile kernel versus backfitting in the partially linear models for longitudinal/clustered data. *Biometrika*, **91**, 251–262.

Huq, N.M. and Cleland, J. (1990) *Bangladesh Fertility Survey 1989*, National Institute of Population Research and Training, Dhaka.

Ishwaran, H. and Gatsonis, C.A. (2000) A general class of hierarchical ordinal regression models with applications to correlated ROC analysis. *Canadian Journal of Statistics*, **28**, 731–750.

Jenss, R.M. and Bayley, N. (1937) A mathematical method for studying the growth of a child. *Human Biology*, **9**, 556–563.

Jones, B. and Kenward, M. (2003) *Design and analysis of cross-over trials*, Chapman & Hall/CRC, London.

Joreskog, K.G. and Sorbom, D. (1979) *Advances in factor analysis and structural equation models*, Abt books, Cambridge, MA.

Kass, R.E. and Steffey, D. (1989) Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models). *Journal of the American Statistical Association*, **84**, 717–726.

Kenward, M. (1998) Selection models for repeated measurements with non random dropout: an illustration of sensitivity. *Statistics in Medicine*, **17**, 2723–2732.

Kenward, M.G. and Roger, J.H. (1997) Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, **53**, 983–997.

Kish, L. (1965) *Survey Sampling*, John Wiley & Sons Inc., New York.

Kish, L. and Frankel, M.R. (1974) Inference from complex samples. *Journal of the Royal Statistical Society, Series B*, **36**, 1–37.

Kreft, I.G., de Leeuw, J. and van der Leeden, R. (1994) Comparing five different statistical packages for hierarchical linear regression: BMDP-5V, GENMOD, HLM, ML3, and VARCL. *American Statistician*, **48**, 324–325.

Kuk, A.Y.C. (1995) Asymptotically unbiased estimation in generalised linear models with random effects. *Journal of the Royal Statistical Society, Series B*, **57**, 395–407.

Laird, N.M. (1988) Missing data in longitudinal studies. *Statistics in Medicine*, **7**, 305–315.

Laird, N.M. and Louis, T.A. (1987) Empirical Bayes confidence intervals based on bootstrap samples. *Journal of the Royal Statistical Society*, **82**, 739–757.

Laird, N.M. and Louis, T.A. (1989) Empirical Bayes confidence intervals for a series of related experiments. *Biometrics*, **45**, 481–495.

Langford, I. and Lewis, T. (1998) Outliers in multilevel data. *Journal of the Royal Statistical Society Series A*, **161**, 121–160.

Langford, I., Leyland, A.H., Rasbash, J. *et al.* (1999) Multilevel modelling of the geographical distributions of diseases. *Journal of the Royal Statistical Society, Series C*, **48**, 253–268.

Larsen, U. and Vaupel, J.W. (1993) Hutterite fecundability by age and parity: strategies for frailty modelling of event histories. *Demography*, **30**, 81–101.

Lawley, D.N. and Maxwell, A.E. (1971) *Factor Analysis as a Statistical Method*, 2nd edn, Butterworth, London.

Leckie, G. and Goldstein, H. (2009) The limitations of using school league tables to inform school choice. *Journal of the Royal Statistical Society, A*, **172**, 835–851.

Leckie, G. and Goldstein, H. (2010) Improved predictions of school performances to inform school choice. Submitted for publication.

Leckie, G., Pillinger, R., Jones, K. *et al.* (2010) Measuring segregation: a multilevel modelling perspective. Submitted for publication.

Lee, Y., Nelder, J.A. and Pawitan, Y. (2006) *Generalized Linear Models with Random Effects*, Chapman & Hall/CRC, Boca Raton.

Lehtonen, R. and Veijanen, A. (1999) *Multilevel Model Assisted Generalised Regression Estimators for Domain Estimation*, International Association of Survey Statisticians Workshop, Riga.

Leiby, B.E., Sammel, M.D., Have, T.R.T. *et al.* (2009) Identification of multivariate responders and non-responders by using Bayesian growth curve latent class models. *Journal of the Royal Statistical Society, Series C*, **58** (4), 505–524.

Lenk, P.J. and DeSarbo, W.S. (2000) Bayesian inference for finite mixtures of generalised linear models with random effects. *Psychometrika*, **65**, 93–119.

Lesaffre, E. and Spiessens, B. (2001) On the effect of the number of quadrature points in a logistic random effects model: an example. *Journal of the Royal Statistical Society, Series C*, **50**, 325–336.

Lewis, T. and Langford, I. (2001) Outliers, robustness and the detection of discrepant data, in *Multilevel Modelling of Health Statistics*, (eds A. Leyland and H. Goldstein), John Wiley & Sons Ltd, Chichester.

Leyland, A. (2001) Spatial Analysis, in *Multilevel Modelling of Health Statistics*, (eds A. Leyland and H. Goldstein), John Wiley & Sons Ltd, Chichester.

Leyland, A. and Goldstein, H., eds (2001) *Multilevel Modelling of Health Statistics*, John Wiley & Sons Ltd, Chichester.

Liang, K. and Zeger, S.L. (1986) Longitudinal data analysis using generalised linear models. *Biometrika*, **73**, 45–51.

Liang, K.Y., Zeger, S.L. and Qaqish, B. (1992) Multivariate regression analyses for categorical data. *Journal of the Royal Statistical Society, Series B*, **54**, 3–40.

Lindley, D.V. and Smith, A.F.M. (1972) Bayes estimates for the linear model. *Journal of the Royal Statistical Society, Series B*, **34**, 1–41.

Lindsey, J.K. (1999) Relationships among sample size, model selection and likelihood regions, and scientifically important differences. *Journal of the Royal Statistical Society, Series D*, **48**, 401–412.

Lindsey, J.K. and Lambert, P. (1998) On the appropriateness of marginal models for repeated measurements in clinical trials. *Statistics in Medicine*, **17**, 447–469.

Lindstrom, M.J. and Bates, D.M. (1990) Nonlinear mixed effects models for repeated measures data, *Biometrics*, **46**, 673–687.

Little, R. (1985) A note about models for selectivity bias. *Econometrica*, **53**, 1469–1474.

Little, R.J.A. (1992) Regression with missing X's: a review. *Journal of the American Statistical Association*, **87**, 1227–1237.

Little, T.D., Schnabel, K.U. and Baumert, J., eds (2000) *Modelling Longitudinal and Multilevel Data*, Lawrence Erlbaum Associates, New Jersey.

Liu, Q. and Pierce, D.A. (1994) A note on Gauss-Hermite quadrature. *Biometrika*, **81**, 13–22.

Longford, N. (1999) Multivariate shrinkage estimation of small area means and proportions. *Journal of the Royal Statistical Society, Series A*, **162**, 227–245.

Longford, N.T. (1987) A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects. *Biometrika*, **74**, 817–827.

Longford, N.T. (1993) *Random Coefficien Models*, Clarendon Press, Oxford.

Longford, N.T. and Muthen, B.O. (1992) Factor analysis for clustered populations. *Psychometrika*, **57**, 581–597.

Lumley, T. (2002) Network meta-analysis for indirect treatment comparisons. *Statistics in Medicine*, **21**, 2313–2324.

Lunn, D.J., Thomas, A., Best, N. *et al.* (2000) WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, **10**, 325–337.

Mathsoft (1999) *S-Plus 2000*. Mathsoft Inc, Seattle.

McCullagh, P. and Nelder, J. (1989) *Generalised Linear Models*, 2nd edition, Chapman and Hall, London.

McCulloch, C.E. (1997) Maximum likelihood algorithms for generalised linear mixed models. *Journal of the American Statistical Association*, **92**, 162–170.

McCulloch, C.E. and Searle, S.R. (2001) *Generalised, Linear and Mixed models*, John Wiley & Sons Inc., New York.

McDonald, R.P. (1985) *Factor Analysis and Related Methods*, Lawrence Erlbaum Associates, New Jersey.

McDonald, R.P. (1993) A general model for two level data with responses missing at random. *Psychometrika*, **58**, 575–585.

McDonald, R.P. (1994) The bilevel reticular action model for path analysis with latent variables. *Sociological Methods and Research*, **22**, 399–413.

McDonald, R.P. and Goldstein, H. (1988) Balanced versus unbalanced designs for linear structural relations in two level data. *British Journal of Mathematical and Statistical Psychology*, **42**, 215–232.

McGrath, K. and Waterton, J. (1986) *British Social Attitudes, 1983–1986*, panel survey technical report, Social and Community Planning Research, London.

Meng, X. and Dyk, D. (1998) Fast EM implementations for mixed effects models. *Journal of the Royal Statistical Society, Series B*, **60**, 559–578.

Miller, R.G. (1974) The Jacknife – a review. *Biometrika*, **61**, 1–15.

Moerbeek, M. and Wong, W.K. (2002) Multiple-objective optimal designs for the hierarchical linear model. *Journal of Officia Statistics*, **18** (2), 291–303.

Moerbeek, M., Van Breukelen, G.J.P. and Berger, M.P.F. (2000) Design issues for experiments in multilevel populations. *Journal of Educational and Behavioural Statistics*, **25**, 271–284.

Moerbeek, M., Van Breukelen, J.P. and Berger, M.P. (2001) Optimal experimental design for multilevel logistic models. *Journal of the Royal Statistical Society, Series D*, **50**, 17–30.

Mok, M. (1995) Sample size requirements for 2-level designs in Educational research. *Multilevel Modelling Newsletter*, **7** (2), 11–15.

Mortimore, P., Sammons, P. Stoll, L. *et al.* (1988) *School Matters*, Open Books, Wells.

Moulton, L.H. and Zeger, S.L. (1989) Analysing repeated measures on generalised linear models via the bootstrap. *Biometrics*, **45**, 381–394.

Muthen, B. and Asparouhov, T. (2009) Multilevel regression mixture analysis. *Journal of the Royal Statistical Society, Series A*, **172**, 639–657.

Muthen, B.O. (1989) Latent variable modelling in heterogeneous populations. *Psychometrika*, **54**, 557–585.

Nagin, D.S. (1999) Analyzing developmental trajectories: a semiparametric, group-based approach. *Psychological Methods*, **4**, 139–157.

Nathan, G. (1983). A simulation comparison of estimators for a regression coefficient under differential non-response. *Communications in Statistics: Theory and Methods*, **11**, 645–659.

Nathan, G. and D. Holt (1980). The effect of survey design on regression analysis. *Journal of the Royal Statistical Society, Series B*, **42**, 377–386.

Ng, E., S.W., Carpenter, J. R., Goldstein, H. *et al.* (2006). Estimation in generalised linear mixed models with binary outcomes by simulated maximum likelihood. *Statistical Modelling*, **6**, 23–42.

Nuttall, D.L., Goldstein, H., Prosser, R. *et al.* (1989) Differential school effectiveness. *International Journal of Educational Research*, **13**, 769–776.

Pan, H. and Goldstein, H. (1997) Multilevel models for longitudinal growth norms. *Statistics in Medicine*, **16**, 2665–2678.

Pan, H. and Goldstein H. (1998) Multilevel repeated measures growth modelling using extended spline functions. *Statistics in Medicine*, **17**, 2755–2770.

Paterson, L. (1991) Socio-economic status and educational attainment: a multidimensional and multilevel study. *Evaluation and Research in Education*, **5**, 97–121.

Peto, R. (1972) Contribution to discussion of paper by D.R. Cox. *Journal of the Royal Statistical Society, Series B*, **34**, 205–207.

Pfeffermann, D. (2001) *Multilevel modelling under informative probability sampling*, 53rd Session of International Statistical Institute, Seoul.

Pfeffermann, D., Skinner, C.J., Holmes, D. *et al.* (1998) Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society, Series B*, **60**, 23–40.

Pinheiro, J.C. and Bates, D.M. (2000) *Mixed Effects Models in S and S-plus*, Springer-Verlag, New York.

Pitt, M., Chan, D. and Kohn, R. (2006) Efficient Bayesian inference for Gaussian copula regression models. *Biometrika*, **93** (3), 537–554.

Plewis, I. (1996) Statistical methods for understanding cognitive growth: a review, a synthesis and an application. *British Journal of Mathematical and Statistical Psychology*, **49**, 25–42.

Plewis, I. (1997) *Statistics in Education*, Edward Arnold, London.

Plewis, I. (1985) *Analysing change*, John Wiley & Sons Ltd, Chichester.

Plewis, I. (1993) Reading Progress, in *A Guide to ML3 for New Users*, (ed. G. Woodhouse), Multilevel Models Project, London.

Pourahmadi, M. (1999) Joint mean-covariance models with applications to longitudinal data: unconstrained parameterisation. *Biometrika*, **86**, 677–690.

Pourahmadi, M. (2000) Maximum likelihood estimation of generalised linear models for multivariate Normal covariance matrix. *Biometrika*, **87**, 425–436.

Rabe-Hesketh, S. and Skrondal, A. (2005) *Multilevel and Longitudinal Modelling Using STATA*, College Station, STATA Press.

Raftery, A.E. (1995) Bayesian model selection in social research. *Sociological Methodology*, **25**, 111–163.

Raftery, A.E. and Lewis, S.M. (1992) How many iterations in the Gibbs sampler?, in *Bayesian Statistics 4*, (eds J.M. Bernardo, J.O. Berger, A.P. Dawid *et al.*). Oxford University Press, Oxford, 765–776.

Rasbash, J. and Goldstein, H. (1994) Efficient analysis of mixed hierarchical and cross classified random structures using a multilevel model. *Journal of Educational and Behavioural Statistics* **19**, 337–350.

Rasbash, J., Steele, F., Browne, W. *et al.* (2009) *A User's Guide to MLwiN Version 2.10*, Centre for Multilevel Modelling, University of Bristol, Bristol.

Raudenbush, S.W. (1995) Maximum likelihood estimation for unbalanced multilevel covariance structure models via the EM algorithm. *British Journal of Mathematical and Statistical Psychology*, **48**, 359–370.

Raudenbush, S.W. (2009) Optimal design software. http://www.wtgrantfoundation.org/ resources/overview/research_tools, accessed June 2010.

Raudenbush, S.W. (1994) Equivalence of Fisher Scoring to Iterative Generalised Least Squares in the Normal case with application to hierarchical linear models. Unpublished.

Raudenbush, S.W. (1993) A crossed random effects model for unbalanced data with applications in cross-sectional and longitudinal research. *Journal of Educational Statistics*, **18**, 321–349.

REALCOM: methodology for realistically complex multilevel modelling (2008). Centre for Multilevel Modelling, Bristol. http://www.cmm.bristol.ac.uk/realcom, accessed June 2010.

Rice, N., Jones, A. and Goldstein, H. (1998) Multilevel models where the random effects are correlated with the fixed predictors: a conditioned iterative generalised least squares estimator (CIGLS), University of York, Centre for Health Economics, York.

Riley, R.D. (2009) Multivariate meta-analysis: the effect of ignoring within-study correlation. *Journal of the Royal Statistical Society, Series A*, **172**, 789–811.

Robinson, W.S. (1950) Ecological correlations and the behaviour of individuals. *American Sociological Review*, **15**, 351–357.

Rodriguez, G. and Goldman, N. (2001) Improved estimation procedures for multilevel models with binary response: a case study. *Journal of the Royal Statistical Society, Series A*, **164**, 339–356.

Rosenbaum, P.R. (1995) *Observational Studies*, Springer-Verlag, New York.

Rosier, M.J. (1987) The second international science study. *Comparative Education Review*, **31**, 106–128.

Rowe, K.J. and Hill, P.W. (1997) Simultaneous estimation of multilevel structural equations to model students' educational progress. Tenth International Congress for School effectiveness and Improvement, Memphis, Tennessee.

Royall, R.M. (1986) Model robust confidence intervals using maximum likelihood estimators. International Statistical Review, **54**, 221–226.

Rubin, D.B. and Thayer, D.T. (1982) EM algorithms for ML factor analysis. *Psychometrika* **47**, 69–76.

Rubin, D.B. (1987) *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons Inc., New York.

Sarndal, C.E., Swensson, B. and Wretman, J.H. (1992) *Model Assisted Survey Sampling*, Springer-Verlag, New York.

Schafer, J.L. (1997) *Analysis of incomplete multivariate data*, Chapman and Hall, London.

Scheuren, F. and Winkler, W.E. (1993). Regression analysis of data files that are computer matched. *Survey Methodology*, **19**, 35–38.

Searle, S.R., Casella, G. and McCulloch, C.E. (1992) *Variance Components*. John Wiley & Sons Inc., New York.

Seltzer, M.H. (1993) Sensitivity analysis for fixed effects in the hierarchical model: a Gibbs sampling approach. *Journal of Educational Statistics*, **18**, 207–236.

Shapiro, A. (1985) Asymptotic distribution of test statistics in the analysis of moment structures under inequality constraints. *Biometrika*, **72**, 133–144

Shi, L. and Chen, G. (2008) Case deletion diagnostics in multilevel models. *Journal of multivariate analysis*, **99**, 1860–1877.

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

Singer, B. and Spilerman, S. (1978) Clustering on the main diagonal in mobility matrices, in *Sociological Methodology 1979*, (ed. K. Schuessler), Jossey-Bass, San Francisco.

Singer, J.D. and Willett, J.B. (2002) *Applied Longitudinal Data Analysis: Modelling Change and Event Occurence.* Oxford University Press, New York.

Skinner, C.J., Holt, D. and Smith, T.M.F (1989) *Analysis of Complex Surveys*, John Wiley & Sons Ltd, Chichester.

Skrondal, A. and Rabe-Hesketh, S. (2004) *Generalized Latent Variable Modelling*, Chapman & Hall /CRC, Boca Raton.

Skrondal, A. and Rabe-Hesketh, S. (2009) Prediction in multilevel generalised linear models. *Journal of the Royal Statistical Society, Series A*, **172** (3), 659–687.

Snijders, T. and Bosker, R. (1999) *Multilevel Analysis*, Sage, London.

Snijders, T.A.B. and Bosker, R.J. (1993) Standard errors and sample sizes for two-level research. *Journal of Educational Statistics*, **18**, 237–259.

Sobel, M.E. (2000) Causal inference in the social sciences. *Journal of the American Statistical Association*, **95**, 647–651.

Speckman, P. (1988) Kernel smoothing in partial linear models. *Journal of the Royal Statistical Society, Series B*, **50**, 413–436.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P. *et al.* (2002) Bayesian measures of model complexity and fit. *Journal of the Royal statistical Society, Series B*, **64**, 583–640.

Steele, F. (2003) Selection effects of source of contraceptive supply in an analysis of discontinuation of contraception: multilevel modelling when random effects are correlated with an explanatory variable. *Journal of the Royal Statistical Society, Series A*, **166**, 407–424.

Steele, F., Diamond, I., Wang, D.L. (1996) The determinants of the duration of contraceptive use in China. A multilevel discrete-hazards modelling approach. *Demography*, **33**, 12–23.

Steele, F., Goldstein, H. and Browne, W. (2004) A general multilevel multistate competing risks model for event history data, with an application to a study of contraceptive use dynamics. *Statistical Modelling*, **4**, 145–159.

Steele, F., Kallis, C., Goldstein, H. *et al.* (2005) The Relationship between childbearing and Transitions from Marriage and cohabitation in Britain. *Demography*, **42**, 647–673.

Tanner, M. and Wong, W.H. (1987) The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, **82**, 528–540.

Touloumi, G., Pocock, S.J., Babiker, A.G. *et al.* (1999). Estimation and comparison of rates of change in longitudinal studies with informative dropouts. *Statistics in Medicine* **18**: 1215–1233.

Turner, R.M., Omar, R.Z., Yang, M. *et al.* (2000) A multilevel model framework for meta analysis of clinical trials with binary outcomes. *Statistics in Medicine*, **19**, 3417–3432.

Turner, R.M., Spiegelhalter, D.J., Smith, G.C.S. *et al.* (2009) Bias modelling in evidence synthesis. *Journal of the Royal Statistical Society, Series A*, **172** (1), 21–47.

Van Buuren, S. (2007) Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, **16** (3), 219–242.

Verbyla, A.P., Cullis, B.R., Kenward, M.G. and Welham, S.J. (1999) The analysis of designed experiments and longitudinal data by using smoothing splines. *Journal of the Royal Statistical Society, Series C*, **48**, 269–312.

Vermunt, J. (2008) Latent class and finite mixture models for multilevel data sets. *Statistical Methods in Medical Research*, **17**, 33–51.

Waclawiw, M.A. and Liang, K. (1993) Prediction of random effects in the generalised linear model. *Journal of the American Statistical Association*, **88**, 171–78.

Waclawiw, M.A. and Liang, K. (1994) Empirical Bayes estimation and inference for the random effects model with binary response. *Statistics in Medicine*, **13**, 541–551.

Wang, Y. (1998) Mixed effects smoothing spline ANOVA. *Journal of the Royal Statistical Society, Series B*, **60**, 159–174.

Wei, L.J., Lin, D.Y. and Weissfeld, L. (1989) Regression analysis of multivariate incomplete failure time data by modelling marginal distributions. *Journal of the American Statistical Association*, **84**, 1065–1073.

Welton, N.J., Ades, A.E., Carlin, J.B. *et al.* (2009) Models for potentially biased evidence in meta analysis using empirically based priors. *Journal of the Royal Statistical Society*, *Series A*, **172** (1), 119–136.

Wolfinger, R. (1993) Laplace's approximation for nonlinear mixed models. *Biometrika*, **80**, 791–795.

Woodhouse, G. (1998) Adjustment for measurement error in multilevel analysis. Institute of Education, University of London, London.

Woodhouse, G. and Goldstein, H. (1989) Educational performance indicators and LEA league tables. *Oxford Review of Education*, **14**, 301–319.

Woodhouse, G., Yang, M., Goldstein, H. *et al.* (1996) Adjusting for measurement error in multilevel analysis. *Journal of the Royal Statistical Society, Series A*, **159**, 201–212.

Wu, H. and Zhang, J. (2002) Local polynomial mixed effects models for longitudinal data. *Journal of the American Statistical Association*, **97**, 883–897.

Yang, M., Goldstein, H. and Heath, A. (2000) Multilevel models for repeated binary outcomes: attitudes and vote over the electoral cycle. *Journal of the Royal Statistical Society, Series A*, **163**, 49–62.

Yang, M., Goldstein, H., Browne, W. and Woodhouse, G. (2001) Multivariate multilevel analyses of examination results. *Journal of the Royal Statistical Society, Series A*, **165**, 137–153.

Zeger, S.L., Liang, K. and Albert, P. (1988) Models for longitudinal data: A generalised estimating equation approach. *Biometrics*, **44**, 1049–1060.

Zeger, S.L. and Karim, M.R. (1991) Generalised linear models with random effects; a Gibbs Sampling approach. *Journal of the American Statistical Association*, **86**, 79–102

Zhou, X., Perkins, A.J. and Hui, S.L. (1999) Comparisons of software packages for generalized linear multilevel models. *American Statistician*, **53**, 282–290.

Author index

Subject index