Rainer Winkelmann

Econometric Analysis of Count Data

Fifth Edition

$$\mathcal{P}(s) = e^{-\lambda} \sum_{k=0}^{\infty} \frac{(\lambda s)^k}{k!}$$

Springer

Econometric Analysis of Count Data

Rainer Winkelmann

Econometric Analysis of Count Data

Fifth edition

Springer

Prof. Dr. Rainer Winkelmann
University of Zurich
Socioeconomic Institute
Zürichbergstr. 14
8032 Zürich
Switzerland
winkelmann@sts.uzh.ch

ISBN 978-3-540-77648-2 e-ISBN 978-3-540-78389-3

DOI 10.1007/978-3-540-78389-3

Library of Congress Control Number: 2008922297

Production: le-tex Jelonek, Schmidt & Vöckler GbR, Leipzig
Cover design: WMX Design GmbH, Heidelberg

Printed on acid-free paper

9 8 7 6 5 4 3 2 1

springer.com

Preface

The "count data" field has further flourished since the previous edition of this book was published in 2003. The development of new methods has not slowed down by any means, and the application of existing ones in applied work has expanded in many areas of social science research. This, in itself, would be reason enough for updating the material in this book, to ensure that it continues to provide a fair representation of the current state of research.

In addition, however, I have seized the opportunity to undertake some major changes to the organization of the book itself. The core material on cross-section models for count data is now presented in four chapters, rather than in two as previously. The first of these four chapters introduces the Poisson regression model, and its estimation by maximum likelihood or pseudo maximum likelihood. The second focuses on unobserved heterogeneity, the third on endogeneity and non-random sample selection.

The fourth chapter provides an extended and unified discussion of zeros in count data models. This topic deserves, in my view, special emphasis, as it relates to aspects of modeling and estimation that are specific to counts, as opposed to general exponential regression models for non-negative dependent variables. Count distributions put positive probability mass on single outcomes, and thus offer a richer set of interesting inferences. "Marginal probability effects" for zeros – at the "extensive margin" – as well as for any positive outcome – at the "intensive margin" – can be computed, in order to trace the response of the entire count distribution to changes in an explanatory variable. The fourth chapter addresses specific methods for flexible modeling and estimation of such distribution responses, relative to the benchmark case of the Poisson distribution.

The organizational changes are accompanied by extensive changes to the presentation of the existing material. Many sections of the book have been entirely re-written, or at least revised to correct for typos and inaccuracies that had slipped through. Hopefully, these changes to presentation and organization have made the book more accessible, and thus more useful also as a reference for graduate level courses on the subject. The list of newly in-

cluded topics includes: Poisson polynomial and double Poisson distribution; the significance of Poisson regression for estimating log-linear models with continuous dependent variable; marginal effects at the extensive margin; additional semi-parametric methods for endogenous regressors; new developments in discrete factor modeling, including a more detailed presentation of the EM algorithm; and copula functions.

I acknowledge my gratitude to those who contributed in various ways, and at various stages, to this book, including Tim Barmby, Kurt Brännäs, Siddharta Chib, Malcolm Faddy, Bill Greene, Edward Greenberg, James Heckman, Robert Jung, Tom Kniesner, Gary King, Nikolai Kolev, Jochen Mayer, Daniel Miles, Andreas Million, Hans van Ophem, Joao Santos Silva, Pravin Trivedi, Frank Windmeijer and Klaus Zimmermann. Large parts of this fifth edition were read by Stefan Boes, Adrian Bruhin and Kevin Staub, and their insights and comments lead to substantial improvements. Part of the revision was completed while I was on leave at the University of California at Los Angeles and at the Center for Economic Studies at the University of Munich. I am grateful for the hospitality experienced at both institutions. In particular, I owe a great debt to doctoral students at UCLA and in Munich, whose feedback to a count data course I was teaching there led, I trust, to substantial improvements in the presentation of the material.

Zürich, January 2008 *Rainer Winkelmann*

Contents

List of Figures

List of Tables

1

Introduction

This book discusses specification and estimation of regression models for non-negative integers, or counts, i.e., dependent variables that take the values $y = 0, 1, 2, \ldots$ without explicit upper limit. Regression analysis, narrowly defined, attempts to explain variation in the conditional mean of y with the help of variation in explanatory variables x. If the mean function is embedded in a probability distribution, one obtains a full conditional probability model of y given x.

Regression and conditional probability models are key tools for the applied researcher who is interested in the relationship between y and x, regardless of whether such relationships are approached from an exploratory or from a confirmatory perspective. If the dependent variable is a count, the econometric all-purpose regression tool, the linear regression model, has a number of serious shortcomings. Hence, more suitable models are required, and the Poisson regression model is the most important count data model.

1.1 Poisson Regression Model

The advantage of the Poisson regression model (PRM) is that it explicitly recognizes the non-negative integer character of the dependent variable. It has two components, first a distributional assumption, and second a specification of the mean parameter as a function of explanatory variables. The Poisson distribution is a one parameter distribution. The parameter, λ, is equal to the mean and the variance, and it must be positive. It is convenient to specify λ as an exponential function of a linear index of the explanatory variables x in order to account for observed heterogeneity: $\lambda = \exp(\beta_1 + \beta_2 x_2 + \ldots + \beta_k x_k)$ or, in vector notation, $\lambda = \exp(x'\beta)$. The exponential form ensures that λ remains positive for all possible combinations of parameters and explanatory variables. Moreover, the systematic effects interact in a multiplicative way, and the coefficients β_j have the interpretation of a partial elasticity of $\mathrm{E}(y|x)$ with respect to (the level of) x_j if the logarithm of x_j is included among the

regressors. The model can be generalized by including non-linear transformations of x_j, for instance a higher order polynomial, among the regressors.

Assuming an independent sample of pairs of observations (y_i, x_i), the parameters of the model can be estimated by maximum likelihood. Although the first-order conditions are non-linear and thus not solvable in closed form, iterative algorithms can be used to find the maximum which is unique as the log-likelihood function is globally concave. Under correct specification, the estimator has all the desirable properties of maximum likelihood estimators, in particular asymptotic efficiency and normality.

The lack of a mean-independent determination of the variance for the Poisson distribution contrast with the flexibility of the two-parameter normal distribution where the variance of the distribution can be adjusted independently of the mean. This feature of the PRM is likely too restrictive. However, Poisson regression is robust: the estimator for β remains consistent even if the variance does not equal the mean (and the true distribution therefore cannot be Poisson) as long as the mean function λ is correctly specified. This robustness mirrors the result for the linear model where OLS is unbiased independently of the second-order moments of the error distribution.

However, it can be inappropriate in other respects. In fact, it is a common finding in applied work using economic count data that certain assumptions of the PRM are systematically rejected by the data. Much of this book is concerned with a unified presentation of the whole variety of count data models that have been developed to date in response to these restrictive features of the PRM.

1.2 Examples

The count model of choice very much depends on the type of available data. In particular, the following questions have to be answered at the outset:

- What is the nature of the count data? Are they univariate or multivariate, are they grouped or censored, what is known about the stochastic process underlying the generation of the data?
- What was the sampling method? Are the data representative of the population, or have they been sampled selectively?

A crude frequency tabulation of the dependent variable can be helpful in selecting an initial model framework. Consider, for instance, the following examples taken from the applied count data literature:

- Kennan (1985) gives the monthly number of contract strikes in U.S. manufacturing. In his analysis, Kennan concentrates on the duration of strikes, rather than on their number per se.
- McCullagh and Nelder (1989) look at the incidence of certain ship damages caused by waves using the data provided by an insurance company. They model the number of incidents regardless of the damage level.

- Zimmermann and Schwalbach (1991) use a data set on the number of patents (stock) of German companies registered at the German Patent Office in 1982. They merge information from the annual reports of the respective companies as well as industry variables.
- Davutyan (1989) studies how the number of failed banks per year in the U.S. for 1947 - 1981 relates to explanatory variables such as a measure of the absolute profitability of the economy, the relative profitability of the banking sector, as well as aggregate borrowing from the Federal Reserve.
- Dionne, Gagné, Gagnon and Vanasse (1997) study the frequency of airline accidents (and incidents) by carrier in Canada on a quarterly basis between 1974 and 1988. Their sample includes approximately 100 Canadian carriers, resulting in around 4000 panel entries. The total number of accidents during the period was 530.
- Winkelmann and Zimmermann (1994) model completed fertility measured by the number of children. Using the German Socio-Economic Panel, they select women aged between 40 and 65 who live in their first marriage. The number of children varies from 0 to 10, the mean is 2.06, and the mode is 2.

Table 1.1. Count Data Frequency Distributions

Counts	Strikes	Ships	Patents	Banks	Airplane	Children
0	-	9	30	-	3498	61
1	12	5	6	-	411	167
2	14	2	7	2	51	297
3	11	1	2	7	3	117
4	9	2	0	4	2	52
5	14	1	3	4	-	14
6	9	2	1	4	-	12
7	4	2	2	1	-	2
8	7	0	0	3	-	1
9	10	0	1	5	-	-
10	6	0	0	3	-	1
> 10	7	11	19	2	-	-
> 100	-	-	20	-	-	-
Observations	103	34	91	35	3965	724
Maximum	18	58	9805	17	4	10
Mean	5.5	10.2	304.6	6.3	0.013	2.1
Variance	13.4	236.5	$1.6*10^6$	11.8	0.015	1.7

The respective empirical frequency distributions of the dependent count variable are given in Tab. 1.1. The six frequency distributions are very indicative of the type of data encountered in applied research. First, the realized

range of observations varies from application to application. In two cases, no zeros are observed, while in other cases, zero is the modal value. Some of the empirical distributions are uni-modal, while others display multiple modes. In most cases, the variance clearly exceeds the mean, while in one case (airlines) it is roughly the same, and in one case (children), the mean is greater than the variance. Second, the structure of the data differs. The three observed types of data are a cross section of individuals, a panel, and a time series. Models for all three types of data are covered in this book.

It should be noted that Tab. 1.1 shows marginal frequencies whereas the focus of this book is on conditional models. Such models account for the influence of covariates in a regression framework. For instance, if the conditional distribution of y given (a non-constant) x is Poisson, the marginal distribution of y cannot be Poisson as well.

1.3 Organization of the Book

Chap. 2 presents probability models for count data. The basic distributions are introduced. They are characterized both through the underlying stochastic process, and through their relationships amongst each other. Most generalizations rely on the tools of mixing and compounding – these techniques are described in some detail. A discussion of hyper-distributions reveals the differences and commonalities between the models. This chapter also draws extensive analogies between probabilistic models for duration data and probabilistic models for count data.

Chap. 3 starts with a detailed exposition of the Poisson regression model, including a comparison with the linear model. Two issues that are of particular relevance for the practitioner are the correct interpretation of the regression coefficients, including inference based on proper standard errors. The basic estimation techniques are discussed, and the properties of the estimators are derived, both under maximum likelihood and pseudo maximum likelihood assumptions. The second part of the chapter is devoted to possible misspecification of the Poisson regression model: its origins, consequences, and how to detect misspecification through appropriate testing procedures.

The bulk of the literature has evolved around three broad types of problems, unobserved heterogeneity, endogeneity, and excess zeros, and these are singled out for special consideration in Chapters 4 – 6, respectively. As far as unobserved heterogeneity is concerned, this leads us from parametric generalizations on one hand (negative binomial model, Poisson-log-normal model), to semi-parametric extensions on the other (series expansions, finite mixtures). Similarly, for endogeneity, instrumental variable estimation via GMM requires minimal moment assumptions. Alternative models are build around a fully specified joint normal distribution for latent errors, and thus, while more efficient if correct, vulnerable to distributional misspecification. Chapter 6 on zeros in count data models presents mostly parametric generalizations, namely

multi-index models, which lead to flexible estimators for marginal probability effects in different parts of the outcome distribution. Quantile regression for counts, a semi-parametric method, is discussed as well.

Chap. 7 is concerned with count data models for multivariate, panel and time series data. This is an area of intensive current research effort, and many of the referred papers are still at a working paper stage. However, a rich class of models is beginning to emerge and the issues are well established: the need for a flexible correlation structure in the multivariate context, and the lack of strictly exogenous regressors in the case of panel data.

Chap. 8 provides an introduction to Bayesian posterior analysis of count data. Again, many of the developments in this area are quite recent. They partly mirror the general revival of applied Bayesian analysis that was triggered by the combined effect of increasing computing power and the development of powerful algorithms for Markov chain Monte Carlo simulation. The potential of this approach is demonstrated, among other things, in an model for highly dimensional panel count data models with correlated random effects.

The final Chap. 9 illustrates the practical use of count data models in a number of applications. Apart from a literature review for applications such as accidents, health economics, demography and marketing, the chapter contains an extended study of the determinants of labor mobility using data from the German Socio-Economic Panel.

2

Probability Models for Count Data

2.1 Introduction

Since probability distributions for counts are not yet entirely standard in the econometric literature, their properties are explored in some detail in this chapter. Special attention is paid to flexible, or 'generalized', count data distributions since they potentially serve as building blocks for improved count data regression models.

Count data frequently arise as outcomes of an underlying *count process* in continuous time. The classical example for a count process is the number of incoming telephone calls at a switchboard during a fixed time interval. Let the random variable $N(t)$, $t > 0$, describe the number of occurrences during the interval $(0, t)$. *Duration analysis* studies the waiting times τ_i, $i = 1, 2, \ldots$, between the $(i-1)$-th and the i-th event. *Count data models*, by contrast, model $N(T)$ for a given T. By studying the relation between the underlying count process, the most prominent being the Poisson process, and the resulting probability models for event counts N, one can acquire a better understanding of the conditions under which a given count distribution is appropriate. For instance, the Poisson process, resulting in the Poisson distribution for the number of counts during a fixed time interval, requires independence and constant probabilities for the occurrence of successive events, an assumption that appears to be quite restrictive in most applications to social sciences or elsewhere. Further results are derived in this chapter.

2.2 Poisson Distribution

2.2.1 Definitions and Properties

Let X be a random variable with a discrete distribution that is defined over $\mathbb{N} \cup \{0\} = \{0, 1, 2, \ldots\}$. X has a Poisson distribution with parameter λ, written $X \sim \text{Poisson}(\lambda)$ if and only if the probability function is as follows:

$$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}, \qquad \lambda \in I\!R^+, \; k = 0, 1, 2, \ldots \tag{2.1}$$

The probability generating function of the Poisson distribution is given by

$$\begin{aligned}\mathcal{P}(s) &= \sum_{k=0}^{\infty} s^k P(X = k) = \sum_{k=0}^{\infty} s^k \frac{e^{-\lambda}\lambda^k}{k!} \\ &= e^{-\lambda} \sum_{k=0}^{\infty} \frac{(\lambda s)^k}{k!} = e^{-\lambda+\lambda s}\end{aligned} \tag{2.2}$$

(See Appendix A for definition and properties of the probability generating function). Conversely, the Poisson probability function is obtained as

$$P(X = k) = (k!)^{-1} \left. \frac{\mathrm{d}^k \mathcal{P}}{\mathrm{d}s^k} \right|_{s=0} = \frac{e^{-\lambda}\lambda^k}{k!} \tag{2.3}$$

The Poisson distribution has expected value

$$\mathrm{E}(X) = \mathcal{P}'(1) = \lambda \tag{2.4}$$

and variance

$$\begin{aligned}\mathrm{Var}(X) &= \mathcal{P}''(1) + \mathcal{P}'(1) - [\mathcal{P}'(1)]^2 = \lambda^2 + \lambda - \lambda^2 \\ &= \lambda\end{aligned} \tag{2.5}$$

Alternatively, the expected value can be derived directly using the probability function:

$$\mathrm{E}(X) = \sum_{k=0}^{\infty} k \frac{e^{-\lambda}\lambda^k}{k!} = \sum_{k=1}^{\infty} k \frac{e^{-\lambda}\lambda^k}{k!} = \lambda \sum_{k=1}^{\infty} \frac{e^{-\lambda}\lambda^{(k-1)}}{(k-1)!} = \lambda \tag{2.6}$$

The equality of mean and variance is characteristic of the Poisson distribution. It plays a crucial role in the further discussion and will be referred to as *equidispersion*. Departures from equidispersion can be either *overdispersion* (variance is greater than the mean) or *underdispersion* (variance is smaller than the mean). In contrast to other multi-parameter distributions, such as the normal distribution, a violation of the variance assumption is sufficient for a violation of the Poisson assumption.

Some Further Properties of the Poisson Distribution

1. The ratio of recursive probabilities can be written as:

$$\frac{p_k}{p_{k-1}} = \frac{\lambda}{k}. \tag{2.7}$$

Thus, probabilities are strictly decreasing for $0 < \lambda < 1$ and the mode is 0; for $\lambda > 1$, the probabilities are increasing for $k \leq \mathrm{int}[\lambda]$ and then decreasing. The distribution is uni-modal if λ is not an integer and the

mode is given by int[λ]. If λ is an integer, the distribution is bi-modal with modes at λ and $\lambda - 1$.

2. Taking the first derivative of the Poisson probability function with respect to the parameter λ, we obtain

$$\frac{dp_k}{d\lambda} = \frac{e^{-\lambda} \cdot (-1)\lambda^k}{k!} + \frac{e^{-\lambda} k \lambda^{k-1}}{k!} = p_k \left(\frac{k}{\lambda} - 1 \right) \tag{2.8}$$

Therefore, the probabilities p_k decrease with an increase in λ (i.e., with an increase in the expected value) for $k < \lambda$. Thereafter, for $k > \lambda$, the probabilities p_k increase with an increase in λ.

3. Consider the dichotomous outcomes $P(X = 0)$ and $P(X > 0)$. The probabilities are given by

$$p_0 = e^{-\lambda}$$

and

$$p_+ = 1 - e^{-\lambda},$$

respectively. These expressions coincide with the cumulative and complementary cumulative density functions of the exponential distribution. The intrinsic relation between the Poisson distribution and the exponential distribution is explored in section (2.2.6).

Sums of Poisson Random Variables

Assume that $X \sim \text{Poisson}(\lambda)$ and $Y \sim \text{Poisson}(\mu)$, λ, $\mu \in \mathbb{R}^+$, and that X and Y are independent. The random variable $Z = X + Y$ is Poisson distributed $Po(\lambda + \mu)$. This result follows directly from the definition of probability generating functions, whereby, under independence, $\text{E}(s^{X+Y}) = \text{E}(s^X)\text{E}(s^Y)$. Further,

$$\begin{aligned} \mathcal{P}^{(Z)} &= \text{E}(s^{X+Y}) \\ &= e^{-(\lambda+\mu)+(\lambda+\mu)s} \end{aligned} \tag{2.9}$$

which is exactly the probability generating function of a Poisson distributed random variable with parameter $(\lambda + \mu)$. Hence, $Z \sim \text{Poisson}(\lambda + \mu)$.

Alternatively, from first principles,

$$\begin{aligned} P(Z = k) &= \sum_{i=0}^{k} P(X = k - i) P(Y = i) \\ &= \sum_{i=0}^{k} \frac{e^{-\lambda} \lambda^{k-i}}{(k-i)!} \frac{e^{-\mu} \mu^i}{i!} \end{aligned}$$

$$= \frac{e^{-\lambda-\mu}}{k!} \sum_{i=0}^{k} \frac{k!}{(k-i)!i!} \lambda^{k-i} \mu^i$$
$$= \frac{e^{-\lambda-\mu}(\lambda+\mu)^k}{k!} \tag{2.10}$$

where the last equality follows from the definition of binomial coefficients.

Linear Transformations

The Poisson distribution is not closed under linear transformations, since a linear transformations on the sample space do not generate again a Poisson distribution with a different value of the parameter λ.

Let $Y = a + bX$ with $X \sim \text{Poisson}(\lambda)$ and a, b arbitrary constants. For Y to be Poisson distributed, it must be true that $\text{E}(Y) = a + b\lambda = \text{Var}(Y) = b^2\lambda$ for any $\lambda > 0$. But the equality holds if and only if $a = 0$ and $b = 0$ or $b = 1$. Thus, Y does not have a Poisson distribution for arbitrary values of a and b.

Shifted Poisson Distribution

The distribution of $Y = a + bX$ for $b = 1$ is sometimes referred to as "shifted" or "displaced" Poisson distribution with probability function

$$P(X = k) = \frac{e^{-\lambda}\lambda^{(k-a)}}{(k-a)!}, \qquad k = a, a+1, a+2, \ldots \tag{2.11}$$

where a generally is taken to be an integer variable, although this is not necessary. For $a > 0$, such a distribution is characterized by underdispersion (see also Chap. 5.1.1).

It can be shown that within a large class of distributions, only the normal distribution is preserved under both location and scale transformation (see Hinkley and Reid, 1991).

2.2.2 Genesis of the Poisson Distribution

In most applications the Poisson distribution is used to model the number of events that occur over a specific time period (such as the number of telephone calls arriving at a switchboard operator during a given hour, the annual number of visits to a doctor, etc.). It is thus of interest to study how the Poisson distribution is related to the intertemporal distribution of events. The next section introduces the general concept needed for the analysis of this issue, the *stochastic process*. The subsequent sections present a number of underlying stochastic models that each give rise to a Poisson distribution for the number of events during the fixed time interval.

The first model is the *Poisson process* in continuous time. The second model introduces the Poisson distribution as a limiting form of a discrete time

stochastic process. Finally, the Poisson distribution arises from independently and identically exponentially distributed interarrival times between events. All three derivations require as their main assumption that events occur completely randomly over time. The underlying randomness is the hallmark of the Poisson distribution.

2.2.3 Poisson Process

The Poisson process is a special case of a count process which, in turn, is a special case of a stochastic process. Hence, some general definitions will be introduced first, before the properties of the Poisson process are presented.

A stochastic process $\{X(t), t \in T\}$ is a collection of random variables (on some probability space) indexed by time.

$X(t)$ is a random variable that marks the occurrence of an event at time t. The underlying experiment itself remains unformalized and the definitions and arguments are framed exclusively in terms of the $X(t)$. If the index set T is an interval on the real line, the stochastic process is said to be a *continuous time* stochastic process. If the cardinal number of T is equal to the cardinal number of $\mathbb{N}$, it is called a *discrete time* stochastic process.

A stochastic process $\{N(t), t \geq 0\}$ is said to be a count process if $N(t)$ represents the total number of events that have occurred before t.

The following properties hold:

1. $N(t) \geq 0$
2. $N(t)$ is integer valued
3. For $s < t$, $N(s) \leq N(t)$
4. For $s < t$, $N(t) - N(s)$ gives the number of events that occurred in the interval (s, t)

A count process is called stationary if the distribution of the number of events in any time interval depends only on the length of the interval:

$$(\forall s > 0) \qquad N(t_2 + s) - N(t_1 + s) \overset{i.d.}{\sim} N(t_2) - N(t_1)$$

A count process has independent increments if the numbers of events which occur in disjoint time intervals are independent.

The Poisson process is a continuous time count process with stationary and independent increments. In other words, it assumes that the occurrence of a random event at a particular moment is independent of time and of the number of events that have already taken place. Let $N(t, t + \Delta)$ be the number of events that occurred between t and $t + \Delta$, $t > 0, \Delta > 0$. The two basic assumptions of the Poisson process can be formalized as follows:

a) The probability that an event will occur during the interval $(t, t + \Delta)$ is stochastically independent of the number of events occurring before t.

b) The probabilities of one and zero occurrences, respectively, during the interval $(t, t+\Delta)$ are given by:

$$P\{N(t,t+\Delta)=1\} = \lambda\Delta + o(\Delta) \tag{2.12}$$

$$P\{N(t,t+\Delta)=0\} = 1-\lambda\Delta + o(\Delta) \tag{2.13}$$

where $o(\Delta)$ represents any function of Δ which tends to 0 faster than Δ, i.e., any function such that $[o(\Delta)/\Delta] \to 0$ as $\Delta \to 0$.

It follows that the probability of an occurrence is proportional to the length of the interval and the proportionality factor is a constant independent of t. Further,

$$\begin{aligned} P\{N(t,t+\Delta)>1\} &= 1 - P\{N(t,t+\Delta)=0\} \\ &\quad -P\{N(t,t+\Delta)=1\} \\ &= o(\Delta)\ . \end{aligned} \tag{2.14}$$

In a sufficiently short interval, the probability of two or more events occurring approaches zero.

Assumptions a) and b) can be restated by saying that the increments of a Poisson process are *independent* and *stationary*: $N(t,t+\Delta)$ and $N(s,s+\Delta)$ are independent for disjoint intervals $(t,t+\Delta)$ and $(s,s+\Delta)$, and $P\{N(t,t+\Delta)=k\}$ is independent of t.

Let $p_k(t+\Delta) = P\{N(0,t+\Delta)=k\}$ denote the probability that k events occurred before $(t+\Delta)$. The outcome $\{N(0,t+\Delta)=k\}$ can be obtained in $k+1$ mutually exclusive ways:

$$\begin{aligned} &\{N(0,t)=k\} \text{ and } \{N(t,t+\Delta)=0\}\ \text{, or} \\ &\{N(0,t)=k-1\} \text{ and } \{N(t,t+\Delta)=1\}\ \text{, or} \\ &\qquad\qquad\vdots \\ &\{N(0,t)=0\} \text{ and } \{N(t,t+\Delta)=k\}\ . \end{aligned}$$

By assumption of independence, the probability of each of the above outcomes equals the product of the single probabilities of its two constituent parts. For example,

$$P[\{N(0,t)=k\} \text{ and } \{N(t,t+\Delta)=0\}] = p_k(t)(1-\lambda\Delta) \tag{2.15}$$

Similarly,

$$P[\{N(0,t)=k-1\} \text{ and } \{N(t,t+\Delta)=1\}] = p_{k-1}(t)\lambda\Delta \tag{2.16}$$

Furthermore, since the outcome "two or more events" has probability zero we get that

$$P[\{N(0,t)=k-j\} \text{ and } \{N(t,t+\Delta)=j\}] = 0$$

for $j \geq 2$. Finally, the two outcomes (2.15) and (2.16) are disjoint, and the probability of their union is therefore given by the sum of their probabilities. Putting everything together, we obtain:

$$p_k(t+\Delta) = p_k(t)(1-\lambda\Delta) + p_{k-1}(t)\lambda\Delta + o(\Delta) \tag{2.17}$$

i.e.

$$\frac{p_k(t+\Delta) - p_k(t)}{\Delta} = -\lambda(p_k(t) - p_{k-1}(t)) + o(\Delta)\ . \tag{2.18}$$

Taking limits for $\Delta \to 0$:

$$\frac{dp_k(t)}{dt} = -\lambda(p_k(t) - p_{k-1}(t)) \tag{2.19}$$

and similarly that

$$\frac{dp_0(t)}{dt} = -\lambda p_0(t) \tag{2.20}$$

The differential equation (2.20) can be solved using the initial condition $p_0(0) = 1$ to obtain

$$p_0(t) = \exp(-\lambda t)$$

Setting $k = 1$ in (2.19) and multiplying through by $\exp(\lambda t)$, we obtain

$$\exp(\lambda t)\frac{dp_1(t)}{dt} + \lambda\exp(\lambda t)p_1(t) = \frac{d}{dt}[\exp(\lambda t)p_1(t)] = \lambda$$

with solution

$$p_1(t) = \lambda t\exp(-\lambda t)$$

Repeated applications of the same procedure for $k = 2, 3, \ldots$ yields the Poisson probability distribution. Alternatively, one can derive directly the probability generating function of the Poisson distribution:

$$\begin{aligned}\frac{\mathrm{d}\mathcal{P}(s;t)}{\mathrm{d}t} &= \frac{\mathrm{d}\sum_{k=0}^{\infty} p_k(t)s^k}{\mathrm{d}t} \\ &= \sum_{k=0}^{\infty}[-\lambda p_k(t) + \lambda p_{k-1}(t)]s^k \qquad (2.21) \\ &= -\lambda\sum_{k=0}^{\infty} p_k(t)s^k + \lambda s\sum_{k=1}^{\infty} p_{k-1}(t)s^{k-1} \qquad (2.22) \\ &= (-\lambda + \lambda s)\mathcal{P}(s;t)\end{aligned}$$

where it is understood that $p_{-1} = 0$. This first order differential equation has solution

$$\mathcal{P}(s;t) = \exp[(-\lambda + \lambda s)t] \tag{2.23}$$

The length of the interval can be normalized to unity, which gives the probability generating function of the standard Poisson distribution.

2.2.4 Generalizations of the Poisson Process

Non-stationarity

A first generalization is to replace the constant λ in (2.12) by a time-dependent variable $\lambda(t)$:

$$P\{N(t, t+\Delta) = 1\} = \lambda(t)\Delta + o(\Delta) \ . \tag{2.24}$$

Define the integrated intensity $\Lambda(t) = \int_0^t \lambda(s)ds$. It can be shown that

$$P\{N(t) = k\} = \frac{e^{-\Lambda(t)}\Lambda(t)^k}{k!} \ . \tag{2.25}$$

$N(t)$ has a Poisson distribution function with mean $\Lambda(t)$. Hence, this generalization does not affect the form of the distribution.

Dependence

In order to explicitly introduce path dependence, it is helpful to rewrite the basic equation defining the Poisson process (2.12) in terms of the conditional probability

$$P\{N(0, t+\Delta) = k+1 | N(0,t) = k\} = \lambda\Delta + o(\Delta)$$

One generalization is to allow the rate λ to depend on the current number of events, in which case we can write

$$P\{N(0, t+\Delta) = k+1 | N(0,t) = k\} = \lambda_k\Delta + o(\Delta)$$

A process of this kind is known in the literature on stochastic processes as a *pure birth* process. The current intensity now depends on the history of the process in a way that, in econometric terminology, is referred to as "occurrence dependence". In this case, N is not Poisson distributed.

There is a vast literature on birth processes. However, much of it is barely integrated into the count data literature. An exception is Faddy (1997), who uses properties of the pure birth process in order to develop generalized count data distributions. This framework can also be used to give a simple re-interpretation of over- and underdispersion. For instance, if $\lambda_0 < \lambda_1 < \lambda_2 < \ldots$ ("positive occurrence dependence") the count N can be shown to be overdispersed relative to the Poisson distribution. Similarly, if $\lambda_0 > \lambda_1 > \lambda_2 > \ldots$ ("negative occurrence dependence") the count N is underdispersed relative to the Poisson distribution. In order to derive parametric distributions based on birth processes, one needs to specify a functional relationship between λ_k and k. For instance, it can be shown that a pure birth process gives rise to a negative binomial distribution if this function is linear, i.e., for $\lambda_k = \alpha + \beta k$. These results and extensions are presented in greater detail in Chap. 2.5.3.

2.2.5 Poisson Distribution as a Binomial Limit

Consider an experiment all outcomes of which can be unambiguously classified as either success (S) or failure (F). For example, in tossing a coin, we may call *head* a success and *tail* a failure. Alternatively, drawing from an urn that contains only red and blue balls, we may call *red* a success and *blue* a failure. In general, the occurrence of an event is a success and the non-occurrence is a failure. Let the probability of a success be denoted by p. Then $0 < p < 1$ and the probability of a failure is given by $q = 1 - p$.

Now suppose that the experiment is repeated a certain number of times, say n times. Since each experiment results in either an F or an S, repeating the experiment produces a series of S's and F's. Thus, in three drawings from an urn, the result red, blue, red, in that order, may be denoted by SFS. The order may represent discrete time. Thus, the first experiment is made at time $t = 1$, the second at time $t = 2$, and the third at time $t = 3$. Thereby, the sequence of outcomes can be interpreted as a discrete time stochastic process. The urn drawing sequence with replacement is the classical example of an independent and stationary discrete time process: The outcomes of experiments at different points in time are independent, and the probability p of a success is constant over time and equal to the proportion of red balls in the urn. In this situation, all permutations of the sequence have the same probability.

Define a variable X as the total number of successes obtained in n repetitions of the experiment. X is called a count variable and n constitutes an upper bound for the number of counts. Under the assumptions of independence and stationarity, X has a binomial distribution function with probability generating function

$$\mathcal{P}(s) = [q + ps]^n \tag{2.26}$$

The binomial distribution and its properties are discussed in Chap. 2.3.2 in greater detail.

Up to this point, n was interpreted as the number of repetitions of a given experiment. To explicitly introduce a time dimension, consider a fixed time interval $(0, T)$ and divide it into n intervals of equal length. p is now the probability of success within the interval. What happens if the number of intervals increases beyond any bound while T is kept constant? A possible assumption is that the probability of a success is proportional to the length of the interval. The length of the interval is given by T/n, where T can be normalized without loss of generality to 1. Denote the proportionality factor by λ. Then $p_n = \lambda/n$, i.e., $p_n n = \lambda$, a given constant. Moreover, let $q_n = 1 - \lambda/n$. Substituting these expressions for P_n and q_n into (2.26) and taking limits, we obtain

$$\begin{aligned} \lim_{n\to\infty} \mathcal{P}(s) &= \lim_{n\to\infty} \left[1 - \frac{\lambda}{n} + \frac{\lambda}{n}s\right]^n \\ &= \lim_{n\to\infty} \left[1 + \frac{\lambda(s-1)}{n}\right]^n \end{aligned} \tag{2.27}$$

$$= e^{\lambda(s-1)}$$

But (2.27) is precisely the probability generating function of the Poisson distribution. Dividing the fixed time period into increasingly shorter intervals, the binomial distribution converges to the Poisson distribution. This result is known in the literature as 'Poisson's theorem' (See Feller, 1968, Johnson and Kotz, 1969). The upper limit for the number of counts implicit in a binomial distribution disappears, and the sample space for the event counts approaches $\mathbb{N}_0$. Also note that in the limit the variance and expectation of the binomial (if they exist) are identical:

$$\begin{aligned} \lim_{n\to\infty} \mathrm{Var}(X) &= \lim_{n\to\infty} [np(1-p)] \\ &= np \\ &= \mathrm{E}(X) \end{aligned} \tag{2.28}$$

As for the Poisson process, this discrete time stochastic process assumed independence and stationarity (i.e., "randomness") of the successive Bernoulli trials.

2.2.6 Exponential Interarrival Times

The durations separating the arrival dates of events are called *waiting times* or *interarrival times*. Let τ_i be the waiting time between the $(i-1)$-th and the i-th event. It follows that the arrival date of the k-th event is given by $\vartheta_k = \sum_{i=1}^{k} \tau_i$, $k = 1, 2, \ldots$ Let $N(T)$ represent the total number of events that have occurred between 0 and T. Following the definitions of Chap. 2.2.3, $\{N(T), T > 0\}$ is a count process, while for fixed T, $N(T)$ is a count variable. The stochastic properties of the count process (and thus of the count) are fully determined once the joint distribution function of the waiting times τ_i, $i \geq 1$, is known. In particular it holds that the probability that at most $k-1$ events occurred before T equals the probability that the arrival time of the k-th event is greater than T:

$$P(N(T) < k) = P(\vartheta_k > T) \tag{2.29}$$

Moreover

$$\begin{aligned} P(N(T) = k) &= P(N(T) < k+1) - P(N(T) < k) \\ &= P(\vartheta_{k+1} > T) - P(\vartheta_k > T) \\ &= F_k(T) - F_{k+1}(T) \end{aligned} \tag{2.30}$$

where F_k is the cumulative density function of ϑ_k and it is understood that $F_0(T) = 1$.

Equation (2.30) fully characterizes the relationship between event counts and durations. In general, $F_k(T)$ is a complicated convolution of the underlying densities of τ_i, which makes it analytically intractable. However, a great simplification arises if τ_i are identically and independently distributed with

a common distribution. The process is then in the form of a renewal process (Cox, 1962), see also Chapter 2.7.2. In particular, assume that $\{\tau_1, \tau_2, \ldots\}$ are independently and identically exponentially distributed variables, all with density function

$$f(\tau) = \lambda e^{-\lambda\tau} \tag{2.31}$$

In order to establish the distribution function of $N(T)$ using (2.30) one first needs to derive the cumulative density function of $\vartheta_k = \sum_{i=1}^{k} \tau_i$. Given the assumption of independent waiting times, the distribution of this k-fold convolution can be derived using the calculus of Laplace transforms (See Feller, 1971). The Laplace transform $\mathcal{L}(s) = \mathrm{E}(e^{-sX})$ is defined for non-negative random variables. It shares many of the properties of the probability generating function defined for integer-valued random variables. In particular, $\mathcal{L}(s) = \mathcal{P}(e^{-s})$ and the Laplace transform of a sum of independent variables equals the product of the Laplace transforms.

The Laplace transform of the exponential distribution is given by

$$\mathcal{L}_\tau(s) = \int_0^\infty e^{-s\tau} dF(\tau) = (1 + s/\lambda)^{-1} \tag{2.32}$$

Under independence

$$\mathcal{L}_\vartheta(s) = [\mathcal{L}_\tau(s)]^k = (1 + s/\lambda)^{-k} \tag{2.33}$$

But (2.33) is the Laplace transform of the *Erlang distribution* with parameters λ and k. The Erlang distribution is a special case of a gamma distribution, with Laplace transform $\mathcal{L}_\vartheta(s) = (1+s/\lambda)^{-\alpha}$ that arises if $\alpha = k$ is an integer, as it is in the present case. For integer k, the cumulative density $F_k(T)$ may be written as (Abramowitz and Stegun, 1968, p. 262; Feller, 1971, p. 11):

$$F_k(T) = 1 - e^{-\lambda T}\left(1 + \lambda T + \frac{(\lambda T)^2}{2!} + \ldots + \frac{(\lambda T)^{k-1}}{(k-1)!}\right) \tag{2.34}$$

Therefore,

$$P(N = k) = F_k(1) - F_{k+1}(1) = \frac{e^{-\lambda}\lambda^k}{k!} \tag{2.35}$$

We conclude that the Poisson distribution arises if the interarrival times are independently exponentially distributed; it requires both independence within the spell (i.e., no duration dependence) and independence between spells (i.e., no occurrence dependence).

2.2.7 Non-Poissonness

Clearly, the Poisson distribution requires strong independence assumptions with regard to the underlying stochastic process, and any violation of these assumptions in general invalidates the Poisson distribution. It will be shown

how occurrence dependence or duration dependence can be modeled, and how both phenomena lead to count data distributions other than the Poisson.

Following Johnson and Kotz (1969, Chap. 9) and Heckman (1981), consider again the urn model that was introduced in Chap. 2.2.5. The urn has a red balls and b blue balls where a red ball stands for the occurrence of an event, and a blue ball for non-occurrence. The probability of an event is therefore given by the proportion $a/(a+b)$ of red balls in the urn. The experiment is repeated k consecutive times.

Different urn schemes for a given individual may be characterized by whether or not the composition of the urn changes in consecutive trials. The case of unchanged composition implies independent trials and this case has been treated in Chap. 2.2.5. It leads to a binomial distribution for the number of successes.

Now, assume instead that the composition of the urn is altered over consecutive trials. There exist three different possibilities. First, the composition changes as the consequence of previous success. This situation is referred to as "occurrence dependence". Second, the composition changes as the consequence of previous non-success. This situation is referred to as "duration dependence". Third, and finally, the composition may change for exogenous reasons independently of the previous process. This situation is referred to as "non-stationarity".

The first two situations, where previous outcomes have an influence on the current experiment, are also known as *contagion* in the statistics literature, while the notion of *state dependence* is more common in the econometrics literature (Heckman and Borjas, 1980, Heckman, 1981). *Positive* contagion indicates that the occurrence of an event makes further occurrences more likely. For *negative* contagion, the opposite holds. Both cases lead to a *contagious* distribution for the number of counts, the Poisson distribution being an example for a non-contagious distribution. Contagious distributions have originally been developed for the theory of accident proneness (Bates and Neyman, 1951).

Occurrence Dependence

Occurrence dependence can be formalized as follows (Johnson and Kotz, 1969, p. 229): Initially, there are a red balls and b blue balls in the urn. One ball is drawn at random. If it is a red ball representing a success, it is replaced together with s red balls. If it is a blue ball, the proportion $a/(a+b)$ is unchanged, i.e., the blue ball is replaced. If this procedure is repeated n times and X represents the total number of times a red ball is drawn, then X has a Pòlya-Eggenberger distribution (Johnson and Kotz, 1969, p. 231). If the number of red balls is increased after a success ($s > 0$), then an occurrence increases the probability of further occurrences and the urn model reflects positive contagion. Johnson and Kotz (1969, p. 231) show that the negative binomial distribution is obtained as a limiting form. (The negative binomial

distribution and its properties are discussed in Chap. 2.3.1). For $s = 0$, the model reduces to the binomial model with independent trials. For $s = -1$, the urn scheme corresponds to a drawing without replacement, leading to a hypergeometric distribution. Thus, the hypergeometric distribution is a distribution for negative contagion.

Corresponding results can be obtained for stochastic processes in continuous time (see also Chap. 2.2.4). For instance, assume that

$$P\{N(0, t + \Delta) = k + 1 | N(0, t) = k\} = \lambda_k \Delta + o(\Delta)$$

This equation defines a pure birth process. If λ_k is an increasing function of k, we have positive occurrence dependence. A constant function gives the Poisson case without occurrence dependence. A decreasing function indicates negative occurrence dependence. It can be shown that the negative binomial model erises if λ_k increases linearly in k.

Duration Dependence

In the urn model for occurrence dependence, the composition of the urn was left unchanged when a blue ball, i.e., a failure, occurred. If failures matter, then the outcome of an experiment depends on the time (number of draws) that has elapsed since the last success. This dependence generates "duration dependence". Again, duration dependence can be analyzed either in discrete time as represented by the urn-model or in continuous time using the concept of (continuous) waiting times. The continuous time approach was already introduced in Chap. 2.2.6. Further details are provided in Chap. 2.7.

Non-Stationarity

Finally, the assumptions of the standard model may be violated because the composition of the urn changes over consecutive trials due to exogenous effects while being unaffected by previous trials. This is the case if the underlying process is *nonstationary*. Non-stationarity does not necessarily invalidate the Poisson distribution.

Heterogeneity

A genuine ambiguity of the relationship between the underlying stochastic process and the count data distribution arises if the population is heterogeneous rather than homogeneous, as was assumed so far. With heterogeneity, the probability of an occurrence becomes itself a random variable.

For instance, in reference to the urn model, individuals may possess distinct urns that differ in their composition of red and blue balls. Unobserved heterogeneity can be modeled through a population distribution of urn compositions. For sampling with replacement (i.e., no dependence), the composition

of *individual* urns is kept constant over time and the trials are thus independent at the individual level. Although past events do not truly influence the composition of individual urns, they provide some information on the proportion of red and blue balls in an individual urn. By identifying individuals with a high proportion of red balls, past occurrences do influence (increase) the expected probability of further occurrences *for that individual*. The model is said to display 'spurious' or 'apparent' contagion.

Again, it can be shown that under certain parametric assumptions on the form of the (unobserved) heterogeneity, the negative binomial distribution arises as the limiting distribution. Recall that the negative binomial distribution may also arise as a limiting form of *true* positive contagion. This fact illustrates one of the main dilemmas of count data modeling: The distribution of the (static) random variable for counts cannot identify the underlying structural stochastic process if heterogeneity is present. This result is also expressed in an 'impossibility theorem' by Bates and Neyman (1951): In a cross section on counts it is impossible to distinguish between true and spurious contagion.

2.3 Further Distributions for Count Data

The main alternative to the Poisson distribution is the *negative binomial* distributions. Count data may be negative binomial distributed if they were generated from a contagious process (occurrence dependence, duration dependence) or if the rate, at which events occur, is heterogeneous. The *binomial* distribution also represents counts, namely the number of successes in independent Bernoulli trials with stationary probabilities, but it introduces an upper bound given by the number of trials n. This upper bound distinguishes it from the Poisson and negative binomial distributions. The *continuous parameter binomial* distribution is a modification of the binolial distribution with continuous parameter n. Finally, the *logarithmic* distribution is discussed because of its role as a mixing distribution for the Poisson distribution. Good further references for these distributions and their properties are Feller (1968) and Johnson and Kotz (1969).

2.3.1 Negative Binomial Distribution

A random variable X has a negative binomial distribution with parameters $\alpha \geq 0$ and $\theta \geq 0$, written $X \sim \text{Negbin}(\alpha, \theta)$, if the probability function is given by

$$P(X = k) = \frac{\Gamma(\alpha + k)}{\Gamma(\alpha)\Gamma(k+1)} \left(\frac{1}{1+\theta}\right)^{\alpha} \left(\frac{\theta}{1+\theta}\right)^{k} \quad k = 0, 1, 2, \ldots \tag{2.36}$$

$\Gamma(\cdot)$ denotes the gamma function such that $\Gamma(s) = \int_0^\infty z^{s-1} e^{-z} dz$ for $s > 0$. This two parameter distribution has probability generating function

$$\mathcal{P}(s) = [1 + \theta(1 - s)]^{-\alpha} \tag{2.37}$$

The mean and variance are given by

$$\mathrm{E}(X) = \alpha\theta \tag{2.38}$$

and

$$\mathrm{Var}(X) = \alpha\theta(1 + \theta) = \mathrm{E}(X)(1 + \theta) \tag{2.39}$$

Since $\theta \geq 0$, the variance of the negative binomial distribution generally exceeds its mean (" overdispersion"). The overdispersion vanishes for $\theta \to 0$.

The negative binomial distribution comes in various parameterizations. From an econometric point of view, the following considerations apply. In order to be able to use the negative binomial distribution for regression analysis the first step is to convert the model into a mean parameterization, say

$$\lambda = \alpha\theta \tag{2.40}$$

where λ is the expected value. Inspection of (2.40) shows that there are two simple ways of doing this.

1. $\alpha = \lambda/\theta$. In this case, the variance function takes the form

   $$\mathrm{Var}(X) = \lambda(1 + \theta)$$

   Hence, the variance is a linear function of the mean. This model is called "Negbin I" (Cameron and Trivedi, 1986).

2. $\theta = \lambda/\alpha$. In this case, the variance function takes the form

   $$\mathrm{Var}(X) = \lambda + \alpha^{-1}\lambda^2$$

   A negative binomial distribution with quadratic variance function results. This model is called "Negbin II".

The probability functions associated with the two models are as follows:

$$\text{Negbin I:}\quad P(X = k) = \frac{\Gamma(\lambda/\theta + k)}{\Gamma(\lambda/\theta)\Gamma(k + 1)} \left(\frac{1}{1 + \theta}\right)^{\lambda/\theta} \left(\frac{\theta}{1 + \theta}\right)^{k} \tag{2.41}$$

and

$$\text{Negbin II:}\quad P(X = k) = \frac{\Gamma(\alpha + k)}{\Gamma(\alpha)\Gamma(k + 1)} \left(\frac{\alpha}{\alpha + \lambda}\right)^{\alpha} \left(\frac{\lambda}{\alpha + \lambda}\right)^{k} \tag{2.42}$$

Although these two types are the most widely used parameterizations in practice, others are possible. For instance, let

$$\alpha = \sigma^{-2}\lambda^{1-k} \text{ and } \theta = \sigma^2\lambda^k$$

As before, $\mathrm{E}(X) = \lambda$. Substitution of α and θ into (2.39) gives

$$\text{Var}(X) = \lambda(1 + \sigma^2\lambda^k)$$

Thus, for $k = 0$ this parameterization reduces to the negative binomial distribution with linear variance function while for $k = 1$, a quadratic variance function is obtained. Winkelmann and Zimmermann (1995) refer to this model as "Negbin$_k$".

Yet another parameterization is often found in the statistics literature (see e.g. DeGroot, 1986), where in the general expression (2.36), $1/(1 + \theta)$ is replaced by p and $\theta/(1 + \theta)$ is replaced by q. If α is an integer, say n, the distribution is called *Pascal* distribution, and it has the interpretation of a distribution of the number of failures that will occur before exactly n successes have occurred in an infinite sequence of Bernoulli trials with probability of success p. For $n = 1$, this distribution reduces to the *geometric* distribution.

$$P(X = k) = pq^k\,, \qquad k = 0, 1, 2, \ldots \tag{2.43}$$

To summarize, the main advantage of the negative binomial distribution over the Poisson distribution is that the additional parameter introduces substantial flexibility into the modeling of the variance function, and thus heteroskedasticity. In particular, it introduces overdispersion, a more general form of heteroskedasticity than the mean-variance equality implied by the Poisson distribution.

Computational Issues

The presence of the Gamma function in the negative binomial probability function can cause numerical difficulties in computing the probabilities on a computer. For instance, consider the Negbin I formulation where terms such as $\Gamma(\lambda/\theta + k)$ need to be evaluated numerically. According to the GAUSS reference manual (Aptech, 1994), the argument of the gamma function must be less than 169 to prevent numerical overflow. The overflow problem can be avoided when one uses the logarithm of the gamma function (as is usually the case in econometrics applications) where an approximation based on Stirling's formula can be used. But even then, the accuracy of the approximation decreases as the argument of the log-gamma function becomes large. Large arguments arise whenever θ is small and the negative binomial distribution approaches the Poisson distribution.

Fortunately, there is a relatively simple way to avoid this difficulty. In particular, the Gamma function follows the recursive relation $\Gamma(x) = (x - 1)\Gamma(x - 1)$. Thus

$$\begin{aligned}\frac{\Gamma(\alpha + k)}{\Gamma(\alpha)} &= \frac{(\alpha + k - 1)(\alpha + k - 2)\cdots(\alpha + k - k)\Gamma(\alpha)}{\Gamma(\alpha)} \\ &= \prod_{j=1}^{k}(\alpha + j - 1)\end{aligned} \tag{2.44}$$

where it is understood that the product equals one for $k = 0$. By suitable change of index, the product can alternatively be expressed as

$$\frac{\Gamma(\alpha + k)}{\Gamma(\alpha)} = \prod_{j=0}^{k-1} (\alpha + j)$$

or as

$$\frac{\Gamma(\alpha + k)}{\Gamma(\alpha)} = \prod_{j=1}^{k} (\alpha + k - j)$$

Finally, taking logarithms of (2.44) we obtain

$$\ln \Gamma(\alpha + k) - \ln \Gamma(\alpha) = \sum_{j=1}^{k} \ln(\alpha + j - 1) \tag{2.45}$$

Relationship to Other Distributions

The negative binomial distribution nests the Poisson distribution. For $X \sim \text{Negbin}(\alpha, \theta)$, let $\theta \to 0$ and $\alpha \to \infty$ such that $\theta\alpha = \lambda$, a constant. The negative binomial distribution converges to the Poisson distribution with parameter λ.

For a proof, consider the probability generating function of the negative binomial distribution, replace θ by λ/α, and take limits.

$$\begin{aligned} \lim_{\substack{\alpha \to \infty \\ \theta\alpha \to \lambda}} \mathcal{P}(s) &= \lim_{\substack{\alpha \to \infty \\ \theta\alpha \to \lambda}} [1 + \theta(1 - s)]^{-\alpha} \\ &= \lim_{\alpha \to \infty} \left[1 + \frac{\lambda(1 - s)}{\alpha}\right]^{-\alpha} \\ &= e^{-\lambda(1-s)} \end{aligned} \tag{2.46}$$

But this is exactly the probability generating function of a Poisson distribution with parameter λ.

An alternative, and somewhat more cumbersome, derivation of this result can be based directly on the probability distribution function

$$\begin{aligned} \lim_{\substack{\alpha \to \infty \\ \theta\alpha \to \lambda}} P(X = k) &= \lim_{\substack{\alpha \to \infty \\ \theta\alpha \to \lambda}} \frac{\Gamma(\alpha + k)}{\Gamma(\alpha)k!} \left(\frac{1}{1 + \theta}\right)^{\alpha} \left(\frac{\theta}{1 + \theta}\right)^{k} \\ &= \lim_{\alpha \to \infty} \frac{\Gamma(\alpha + k)}{\Gamma(\alpha)k!} \left(\frac{\alpha}{\alpha + \lambda}\right)^{\alpha} \left(\frac{\lambda}{\alpha + \lambda}\right)^{k} \\ &= \lim_{\alpha \to \infty} \left(\prod_{j=1}^{k} \frac{\alpha + j - 1}{\alpha + \lambda}\right) \left(\frac{\alpha}{\alpha + \lambda}\right)^{\alpha} \frac{\lambda^k}{k!} \\ &= \lim_{\alpha \to \infty} \left(\prod_{j=1}^{k} \frac{1 + (j - 1)/\alpha}{1 + \lambda/\alpha}\right) \left(\frac{1}{1 + \lambda/\alpha}\right)^{\alpha} \frac{\lambda^k}{k!} \end{aligned}$$

$$= e^{-\lambda}\frac{\lambda^k}{k!}$$

where use was made of the product expression for the ratio of gamma functions and of the fact that $(\alpha+\lambda)^{-k} = \prod_{j=1}^{k}(\alpha+\lambda)^{-1}$.

Further Characterization of the Negative Binomial Distribution

The negative binomial distribution arises in a number of ways. It was mentioned in Chap. 2.2.7 that it is the limiting distribution of a sequence of non-independent Bernoulli trials. It also arises as a mixture distribution and as a compound distribution. For mixing, assume that $X \sim \text{Poisson}(\lambda)$ and that λ has a gamma distribution. The marginal distribution of X is then the negative binomial distribution. For compounding, assume that a Poisson distribution is compounded by a logarithmic distribution. The compound distribution is then the negative binomial distribution. Derivations of these two results are postponed until Chap. 2.5.1 and Chap. 2.5.2 where the general approaches of mixing and compounding are presented.

Sums of Negative Binomial Random Variables

Assume that X and Y are independently negative binomial distributed with $X \sim \text{Negbin I}(\lambda,\theta)$ and $Y \sim \text{Negbin I}(\mu,\theta)$. It follows that the random variable $Z = X + Y$ is negative binomial distributed $\text{Negbin I}(\lambda+\mu,\theta)$. For a proof, recall that the generic probability generating function of the negative binomial distribution is given by $\mathcal{P}(s) = [1+\theta(1-s)]^{-\alpha}$. In Negbin I parameterization, we obtain

$$\mathcal{P}(s)^{(X)} = [1+\theta(1-s)]^{-\lambda/\theta}$$

and

$$\mathcal{P}(s)^{(Y)} = [1+\theta(1-s)]^{-\mu/\theta}$$

Thus

$$\begin{aligned}\mathcal{P}(s)^{(Z)} &= [1+\theta(1-s)]^{-\lambda/\theta}[1+\theta(1-s)]^{-\mu/\theta} \\ &= [1+\theta(1-s)]^{-(\lambda+\mu)/\theta} \end{aligned} \tag{2.47}$$

Thus, negative binomial distributions of the type specified above are closed under convolution.

This result depends critically on two assumptions: First, the Negbin I specification with linear variance function has to be adopted. Second, X and Y have to share a common variance parameter θ. In other words, the sum of two arbitrarily specified negative binomial distributions is in general **not** negative binomial distributed.

2.3.2 Binomial Distribution

A random variable X has a *binomial distribution* with parameters $n \in \mathbb{N}$, and $p \in (0,1)$, written $X \sim B(n,p)$), if

$$P(X=k) = \binom{n}{k} p^k (1-p)^{n-k} \qquad k = 0, 1, \ldots, n\,. \tag{2.48}$$

The probability generating function is given by

$$\begin{aligned} \mathcal{P}(s) &= \sum_{k=0}^{n} s^k \binom{n}{k} p^k q^{n-k} \\ &= \sum_{k=0}^{n} \binom{n}{k} (ps)^k q^{n-k} \\ &= (q+ps)^n\ . \end{aligned} \tag{2.49}$$

and mean and variance are

$$\mathrm{E}(X) = np$$

and

$$\mathrm{Var}(X) = np(1-p),$$

respectively.

In estimation problems, the binomial parameter n is usually treated as given. Sometimes, however, one might wish to estimate n as a function of data as well. Under maximum likelihood, there are two possibilities. First, one can respect the integer nature of the parameter and maximize by way of a grid search. The resulting estimator won't have the standard properties of a maximum likelihood estimator. Alternatively, one can treat n as a continuous parameter. In this case, derivatives can be taken. Since

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{\Gamma(n+1)}{\Gamma(k+1)\Gamma(n-k+1)}$$

where $\Gamma(\cdot)$ denotes the gamma-function and $\Gamma(n+1) = n!$ if n is an integer, this involves computation of the digamma function. Alternatively, direct differentiation can be based on an approximation of the factorial representation using Stirling's formula

$$k! \approx (2\pi)^{1/2} k^{k+1/2} \exp(-k)\{1 + 1/12k\}$$

In either case, a logical difficulty arises with respect to the possible sample space of the underlying random variable X if n is a continuous non-negative parameter. Consider the following formal definition.

A random variable X has a *continuous parameter binomial* distribution with parameters $\alpha \in \mathbb{R}^+$, and $p \in (0,1)$, written $X \sim CPB(\alpha, p)$, if the nonnegative integer n in equation 2.48 is replaced by a continuous $\alpha \in \mathbb{R}^+$ where $k = 0, 1, \ldots, \tilde{n}$ and

$$\tilde{n} = \begin{cases} \text{int}[\alpha] + 1 & \text{if } \alpha \text{ non-integer} \\ \alpha & \text{if } \alpha \text{ integer} \end{cases}$$

(this is the so-called *ceiling function*, see also Johnson and Kotz, 1969, p.41, King, 1989b). When α is not an integer, the probabilities do not sum to one and the following normalization is used:

$$\tilde{p}_k = \frac{p_k}{\sum_{i=0}^{\tilde{n}} p_i}, \quad k = 0, 1, \ldots, \tilde{n}. \tag{2.50}$$

where

$$p_k = \frac{\Gamma(\alpha+1)}{\Gamma(k+1)\Gamma(\alpha-k+1)} p^k (1-p)^{\alpha-k}$$

However, this formulation has the defect that the expected value is not equal to αp, as the analogy to the binomial distribution would suggest. References that have ignored this point or were at least unclear about it include Guldberg (1931), Johnson and Kotz (1969), and King (1989b). For example, for $0 < \alpha < 1$, there are two possible values for k, 0 or 1, and, using the above definitions,

$$p_0 = \frac{\alpha\Gamma(\alpha)}{\Gamma(1)\Gamma(\alpha+1)} p^0 (1-p)^{\alpha-0} = (1-p)^{\alpha}$$

$$p_1 = \frac{\alpha\Gamma(\alpha)}{\Gamma(2)\Gamma(\alpha)} p^1 (1-p)^{\alpha-1} = \alpha p (1-p)^{\alpha-1}$$

Moreover, with $\tilde{p}_1$ defined as in (2.50),

$$\begin{aligned} \text{E}(X) = \tilde{p}_1 &= \frac{\alpha p (1-p)^{\alpha-1}}{(1-p)^{\alpha} + \alpha p (1-p)^{\alpha-1}} \\ &= \alpha p \left[\frac{1}{1 + (\alpha-1)p} \right] \; > \; \alpha p \end{aligned}$$

The correct computation of the expected value of the continuous parameter binomial distribution for arbitrary α needs to be based on the generic formula

$$\text{E}(X) = \sum_{k=1}^{\tilde{n}} k \tilde{p}_k. \tag{2.51}$$

Winkelmann, Signorino, and King (1995) show that the difference between αp and the correct expected value (2.51) is not large, but it is not zero, and it varies with the two parameters of the CPB. The lack of a simple expression for the expected value somewhat limits the appeal of this distribution for practical work.

2.3.3 Logarithmic Distribution

The random variable X has a logarithmic distribution if (Johnson and Kotz, 1969, p. 166)

$$P(X = k) = \alpha\theta^k/k \qquad k = 1, 2, \ldots,\ 0 < \theta < 1 \tag{2.52}$$

where $\alpha = -[\log(1-\theta)]^{-1}$. The probability generating function is given by

$$\begin{aligned}\mathcal{P}(s) &= \sum_{k=1}^{\infty} s^k \alpha \frac{\theta^k}{k} \\ &= \sum_{k=1}^{\infty} \alpha \frac{(\theta s)^k}{k} \\ &= -\alpha \ln(1-\theta s)\end{aligned} \tag{2.53}$$

where the last equality follows from a Taylor series expansion of $\ln(1-x)$ around 0:

$$\ln(1-x) = -\sum_{k=1}^{\infty} \frac{x^k}{k} \tag{2.54}$$

Alternatively, the probability generating function can be written using the explicit expression of the normalizing constant α as

$$\mathcal{P}(s) = [\log(1-\theta s)]/[\log(1-\theta)] \tag{2.55}$$

The mean and the variance are given by

$$\mathrm{E}(X) = \alpha\theta(1-\theta)^{-1} \tag{2.56}$$

and

$$\mathrm{Var}(X) = \alpha\theta(1-\alpha\theta)(1-\theta)^{-2}\ . \tag{2.57}$$

The distribution displays overdispersion for $0 < \alpha < 1$ (i.e., $\theta > 1 - e^{-1}$) and underdispersion for $\alpha > 1$ (i.e., $\theta < 1 - e^{-1}$).

In contrast to the previous distributions, the sample space of the logarithmic distribution is given by the set of *positive* integers. And in fact, it can be obtained as a limiting distribution of the truncated-at-zero negative binomial distribution (Kocherlakota and Kocherlakota, 1992, p.191). The likely reason for the logarithmic distribution being an ineffective competitor to the Poisson or negative binomial distributions is to be seen in its complicated mean function which factually, though not formally, prohibits the use of the distribution in a regression framework. For instance, Chatfield, Ehrenberg and Goodhardt (1966) use the logarithmic distribution to model the numbers of items of a product purchased by a buyer in a specified period of time, but they do not include covariates, i.e., they specify no regression. However, the logarithmic distribution plays a role as a compounding distribution (see Chap. 2.5.2).

2.3.4 Summary

The main properties of the described distributions for counts are summarized in Tab. 2.1.

Table 2.1. Distributions for Count Data

Distribution	Range	$\mathcal{P}(s)$	E(X)	Var(X)
Poisson	$0, 1, 2, \ldots$	$e^{-\lambda+\lambda s}$	λ	λ
Binomial	$0, 1, \ldots, n$	$(q+ps)^n$	np	$np(1-p)$
Negative Binomial	$0, 1, 2, \ldots$	$[1+\theta(1-s)]^{-\alpha}$	$\alpha\theta$	$\alpha\theta(1+\theta)$
Logarithmic	$1, 2, \ldots$	$-\alpha \ln(1-\theta s)$	$\frac{\alpha\theta}{1-\theta}$	$\frac{\alpha\theta(1-\alpha\theta)}{(1-\theta)^2}$

It is worth emphasizing that the first three distributions display a similar structure. In fact, they are related through various limiting forms that have been discussed in this chapter. The common structure of the distributions can be best captured by considering the following generic probability generating function (Johnson and Kotz, 1969, p. 138):

$$\mathcal{P}(s) = [(1+\omega) - \omega s]^{-m} \tag{2.58}$$

From (2.58) it follows directly that

$$\mathrm{E}(X) = m\omega$$

and

$$\mathrm{Var}(X) = m\omega(1+\omega)$$

The probability generating functions in Tab. 2.1 can be obtained as follows. For the negative binomial model, $\omega > 0$ and $m > 0$; for the binomial, $-1 < \omega < 0$ and $m < 0$. The Poisson distribution is obtained as the limiting intermediate case where $\omega \to 0$ and $m \to 0$ such that $\omega m = \lambda$.

Finally, the following figures compare the shape of the four probability functions for specific parameter values. In all figures, the expected value is set to 3.5 . Fig. 2.1 presents the Poisson distribution, the negative binomial distribution with $\mathrm{Var}(X)/\mathrm{E}(X) = 2$, the binomial distribution with $n = 10$ and the logarithmic distribution ($x \geq 1$). Fig. 2.2 shows the negative binomial distribution for varying degrees of dispersion ($\mathrm{Var}(X)/\mathrm{E}(X) = 1.5$: solid; and $\mathrm{Var}(X)/\mathrm{E}(X) = 3$: shaded).

The figures illustrate the different assumptions on the variance. Taking the Poisson distribution as reference distribution, the binomial distribution is more, and the negative binomial distribution is less concentrated around the mean. The concentration of the negative binomial distribution decreases with

Fig. 2.1. Count Data Distributions ($\mathrm{E}(X) = 3.5$)

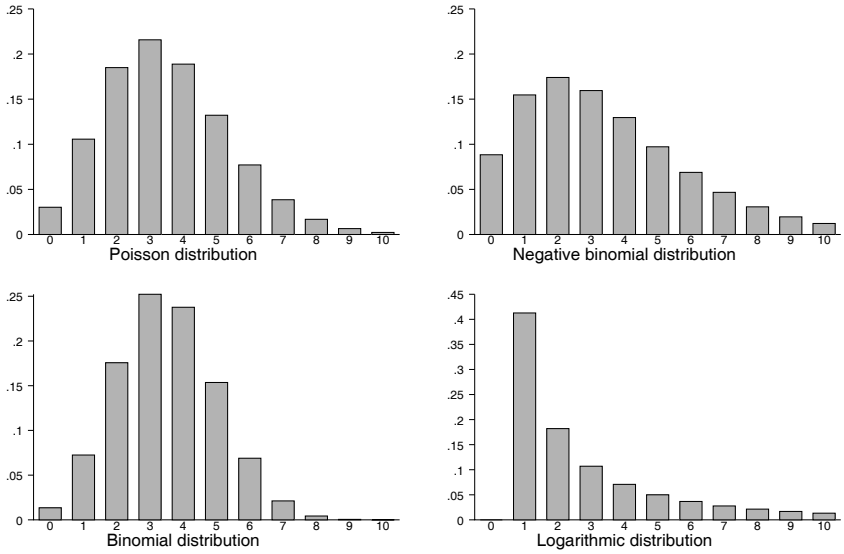
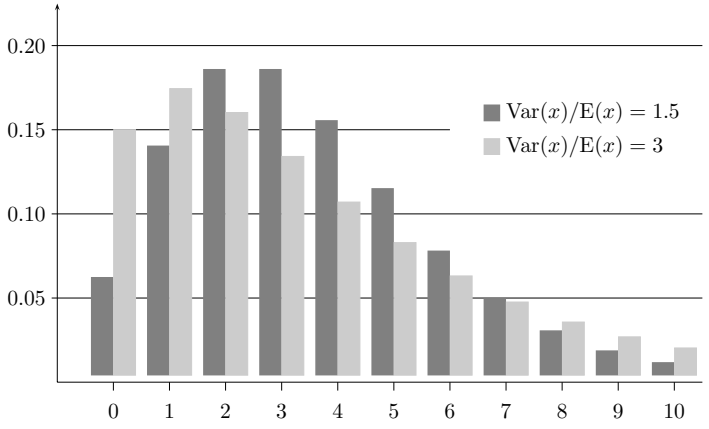


Fig. 2.2. Negative Binomial Distributions with Varying Degrees of Dispersion



increasing variance mean ratio. Another useful comparison is the probability of zeros under the different models. While the probability of obtaining a zero is 1.3 percent under the binomial model, it is 14.6 percent under the negative binomial model with maximal dispersion, the Poisson model providing an intermediate case with 3 percent probability of obtaining a zero (The logarithmic distribution is defined only for positive integers). Finally, it is worth noticing that all distributions are skewed to the left.

2.4 Modified Count Data Distributions

2.4.1 Truncation

If a count data distribution is observable not over the whole range of non-negative integers but rather only for a subset it is said to be truncated. For instance, if observations with zero outcomes are not observed, the distribution is "truncated-at-zero". In this case, "positive count data models" are required (Gurmu, 1991).

Truncated count data can be modeled as a two-part process. The first part consists of an untruncated latent distribution for X^*. The second part consists of a binary indicator variable c. The observed distribution for X is truncated if $c = 0$, and untruncated if $c = 1$. The generic model for truncation is then

$$X = \begin{cases} X^* \\ \text{unobserved} \end{cases} \text{ if } \begin{cases} c = 1 \\ c = 0 \end{cases} \tag{2.59}$$

Further, assume that

$$c = \begin{cases} 1 \text{ if } X^* \in A \\ 0 \text{ if } X^* \notin A \end{cases} \tag{2.60}$$

that is, c is uniquely determined through the latent count variable X^*. The two most commonly encountered situations are:

1. A is the set of positive integers ("truncation at zero").
2. A is the set $\{0, \ldots, a\}$ where a is some positive integer ("truncation from above").

For instance, assume that c is defined as in (2.60) and X^* is Poisson distributed with parameter λ. For $A = \{1, 2, \ldots\}$

$$P(c = 1) = 1 - \exp(-\lambda)$$

and for $A = \{0, 1, \ldots, a\}$

$$P(c = 1) = F(a)$$

where F is the cumulative distribution function of X^*. In general,

$$P(X = k) = \frac{P(X^* = k|c = 1)}{P(c = 1)} \tag{2.61}$$

For the truncated-at-zero Poisson model, we have

$$P(X = k|X > 0) = \frac{\exp(-\lambda)\lambda^k}{k!(1 - \exp(-\lambda))}, \qquad k = 1, 2, 3 \ldots \tag{2.62}$$

with mean

$$\mathrm{E}(X|X > 0) = \frac{\lambda}{1 - \exp(-\lambda)} \tag{2.63}$$

and variance

$$\mathrm{Var}(X|X > 0) = \mathrm{E}(X|X > 0) \left(1 - \frac{\lambda}{\exp(\lambda) - 1}\right). \tag{2.64}$$

Since λ (the mean of the untruncated distribution) is greater than zero, $0 < \exp(-\lambda) < 1$ and the truncated mean is shifted to the right. Moreover, the truncated-at-zero model displays underdispersion since $0 < 1-\lambda(\exp(\lambda)-1) < 1$.

2.4.2 Censoring and Grouping

A count data distribution is said to be censored if it is only partially observable: for a subset of outcomes the distribution is determined only up to an interval of outcomes. The leading example here is right-censoring, where all counts exceeding a certain threshold number k are reported in a category "k or more". For instance, such data are occasionally observed in household survey data (See Merkle and Zimmermann, 1992).

Denote the interval of partial observability by A. Then

$$P(X = k) = p_k \qquad \text{for } k \in I\!N \setminus A$$
$$P(X \in A) = \sum_{k \in A} p_k$$

Censoring can be seen as a special case of grouping. Assume that the set of non-negative integers is partitioned into J mutually exclusive and exhaustive subsets $A_1, \ldots, A_J$, and that each $A_j, j = 1, \ldots, J$ is the set of consecutive integers $\{a_j, a_j + 1, \ldots, a_j + n_j\}$ such that $a_{j+1} = a_j + n_j + 1$ and $a_1 = 0$.

Hence, the set A_j to which a count belongs is known, but not the count itself. The resulting model is defined over the subsets with $P(A_j) = P(X \in A_j)$, where

$$P(X \in A_j) = \sum_{k \in A_j} P(X = k) \tag{2.65}$$

2.4.3 Altered Distributions

For discrete distributions, it is relatively straightforward to select one (or more) specific outcome and increase (or decrease) the probability of that outcome relative to the probability of the underlying model. The only two restrictions are the fundamental requirements for probabilities, namely that they are non-negative and sum up to one. Such a modeling strategy certainly can improve the ability of the probability model to describe actual discrete data. While this approach may appear ad-hoc at first glance, there are situations where adjustments to single probabilities can in fact be justified in terms of underlying structural processes. This idea will be followed up in a later chapter where regression models based on altered count data distributions are presented.

In practice, the most common alteration is to modify the probability of a zero relative to the underlying distribution. The resulting distributions are referred to as "zero-inflated" or "zero-deflated" count data distributions. For instance, the zero-inflated Poisson distribution can be written as

$$P(X = 0) = \omega + (1 - \omega)e^{-\lambda}$$

$$P(X = k) = (1 - \omega)\frac{e^{-\lambda}\lambda^k}{k!} \qquad k = 1, 2, \ldots$$

where $0 < \omega < 1$, or, more compactly, as

$$P(X = k) = \delta_{k=0}\omega + (1 - \omega)\frac{e^{-\lambda}\lambda^k}{k!} \qquad k = 0, 1, 2, \ldots \tag{2.66}$$

where δ is a binary indicator taking the value 1 when $k = 0$ and zero otherwise. "Zero-deflation" would be obtained for $0 > \omega > -(1 - e^{-\lambda})^{-1}$.

Along the same lines one could adjust more than one probability. For instance, consider a zero-and-two inflated model that has been used in a study of completed fertility, i.e., the number of children born (Melkersson and Rooth, 2000): Now

$$P(X = 0) = \omega_0 + (1 - \omega_0 - \omega_2)e^{-\lambda}$$

$$P(X = 2) = \omega_2 + (1 - \omega_0 - \omega_2)\frac{e^{-\lambda}\lambda^2}{2!}$$

$$P(X = k) = (1 - \omega_0 - \omega_2)\frac{e^{-\lambda}\lambda^k}{k!} \qquad k = 1, 3, 4, \ldots$$

Obviously, the expected value and variance of such a distribution is quite different from the expected value and variance of the Poisson distribution.

2.5 Generalizations

Having to choose between one of the probability models introduced in the previous chapters, the researcher might encounter two problems:

- None of the distributions accounts for the full amount of the observed variation in the counts. In particular, the observed frequency distribution might exhibit extraordinary values for some categories, and/or the second moment assumption of the probability model might be too restrictive.
- Although one of the probability models does in fact correspond to the true data generating process, the process is unknown. In other words, the researcher has no *a-priori* information on the true model and he is left with the problem of choosing a particular distribution.

Both issues have been recognized and addressed by the literature leading to the development of so-called 'generalized' or 'modified' probability models. Two types of generalizations can be distinguished. The first type concentrates on additional, possibly more flexible, probability models. Consul (1989) lists generalizations based on the Poisson distribution. They include the "Neyman-type distributions, the Poisson binomial, Poisson negative binomial, Poisson geometric, Poisson Pascal, Poisson rectangular, Poisson log-normal, quasi Poisson, inflated Poisson, mixed Poisson, generalized Poisson, and double Poisson (...)" (Consul, 1989, p.3). Many of these models fall within the class of *compound* or *mixed* Poisson distribution families, which are presented in the next section. Alternatively, more general models have been derived from an application of birth processes.

The second type of generalization addresses the issue of selecting a specific model. Here, a *hyper-model* (or class of distribution families) encompasses several sub-models of interest. Examples are the *Katz* class of distributions and the class of *linear exponential families* which are introduced in Chap. 2.5.4 and Chap. 2.5.6, respectively. Both classes contain the most important distributions for count data – the Poisson, binomial and negative binomial distributions – in the form of either parametric (Katz) or functional restrictions (linear exponential family).

The distinction between the two types of generalizations blurs sometimes, when generalized distributions nest more than one interesting sub-model. In these cases, they can be used either for the benefit of a more flexible model *per se*, or they can serve to discriminate between more restrictive sub-models.

2.5.1 Mixture Distributions

Mixture distributions play an important role in the modeling of counts (but their importance is by no means limited to count data). In general terms, mixtures are defined in the following way: consider various proper distribution functions F^j representing different random variables X^j, $j = 1, 2, \ldots$, and constants a_j with $a_j > 0 \; \forall j$ and $\sum_j a_j = 1$. Then

$$F = \sum_{j=1}^{\infty} a_j F^j \tag{2.67}$$

is a proper distribution function and is called a mixture of the distributions $\{F^j\}$. The component distributions do not have to be defined over the same sample space $\mathbf{S}$. Let $\mathbf{S}_j$ denote the sample space of distribution j and let $\mathbf{S}$ denote the sample space of the mixture distribution F. Then $\mathbf{S}=\bigcup_j \mathbf{S}_j$.

There are various ways by which the general concept of mixing can be given more specific content. For instance, it can be used to give special weights to specific discrete values. For this purpose, one might mix for instance a Poisson distribution with a degenerate distribution putting unity probability mass at one point. The resulting mixture is an "inflated parameter probability" distribution.

Alternatively, F^j might be any parametric distribution function depending on a parameter θ. Moreover, assume that the parameter itself is a random variable with probability function $f(\theta)$. Thus, if the support of θ is discrete, we can write

$$F = \sum_{\theta \in \Theta} f(\theta) F(\theta) \tag{2.68}$$

whereas for continuous support, an integral replaces the summation

$$F = \int_{\theta \in \Theta} f(\theta) F(\theta) d\theta \tag{2.69}$$

Mixtures of this form are commonly expressed in terms of probability functions (rather than distribution functions). In the case of continuous mixing over a discrete probability function, we can write for instance

$$P(X = k) = \int_{\theta \in \Theta} P(X = k|\theta) f(\theta) d\theta \tag{2.70}$$

This last formulation makes it clear that mixing is really a randomization of a distribution parameter. In this scenario, two distinct distributions for X can be distinguished: the *conditional* distribution $P(X = k|\theta)$, and the *marginal* distribution $P(X = k)$. If the marginal distribution of θ is known, then the marginal distribution of X is obtained by integrating the joint distribution of X and θ over θ. An example for such an operation based on a Poisson distribution is given shortly. First, however, it is useful to study the mean and variance of the marginal distribution of X under mixing. Results can be established under very mild assumptions, whereas the derivation of the full marginal distribution requires knowledge of $f(\theta)$, a much stronger requirement.

Mean and Variance of Marginal Distribution

By the law of the iterated expectation

$$\mathrm{E}(X) = \mathrm{E}_\theta[\mathrm{E}(X|\theta)] \tag{2.71}$$

and, using the variance decomposition theorem,

$$\mathrm{Var}(X) = \mathrm{E}_\theta[\mathrm{Var}(X|\theta)] + \mathrm{Var}_\theta[\mathrm{E}(X|\theta)] \tag{2.72}$$

A number of results follow. First, if the conditional distribution of $X|\theta$ is Poisson, then

$$\begin{aligned}\mathrm{E}_\theta[\mathrm{Var}(X|\theta)] &= \mathrm{E}_\theta[\mathrm{E}(X|\theta)] \\ &= \mathrm{E}(X)\end{aligned}$$

where the first equality uses the equi-dispersion property of the Poisson. Therefore,

$$\mathrm{Var}(X) = \mathrm{E}(X) + \mathrm{Var}_\theta[\mathrm{E}(X|\theta)] > \mathrm{E}(X)\,,$$

and mixing introduces overdispersion at the marginal level. An immediate consequence of this result in the context of a multivariate random variable is that it cannot be the case that both marginal and conditional distributions are of the Poisson type.

If we specify the mean and variance of the distribution of θ as $\mathrm{E}(\theta) = \lambda$ and $\mathrm{Var}(\theta) = \sigma_\theta^2$, then an application of (2.71) and (2.72) yields that $\mathrm{E}(X) = \lambda$ and $\mathrm{Var}(X) = \lambda + \sigma_\theta^2\lambda^2$. As the reader may recall, these expressions are equal to the mean and variance of the Negbin II model introduced in Chap. 2.3.1. This is not a coincidence, since the Negbin II model can be derived from mixing a Poisson distribution with a gamma distribution. However, it should be noted that the *semi-parametric* result derived from applying the law of iterative expectations is more general as it does not depend on the full density $f(\theta)$ but only on its first two moments.

Example for a Fully Parametric Mixture Model: Poisson-gamma

The leading example of a fully parametric mixture model for count data is the Poisson-Gamma mixture. Assume that $X \sim \mathrm{Poisson}(\theta)$ where θ is gamma distributed with density function

$$f(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\theta\beta} \tag{2.73}$$

Mean and variance of the gamma distribution are given by $\mathrm{E}(\theta) = \alpha/\beta$ and $\mathrm{Var}(\theta) = \alpha/\beta^2$, respectively. Under the re-parameterization $\beta = \alpha/\lambda$, we obtain the desired specification where $\mathrm{E}(\theta) = \lambda$ and $\mathrm{Var}(\theta) = \alpha^{-1}\lambda^2$. Moreover, integration of $P(X = k, \theta) = P(X = k|\theta)f(\theta)$ over θ yields

$$\begin{aligned}P(X = k) &= \int_0^\infty \frac{e^{-\theta}\theta^k}{k!}\frac{(\alpha/\lambda)^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\frac{\theta\alpha}{\lambda}}d\theta \\ &= \frac{\alpha^\alpha}{\lambda^\alpha k!\Gamma(\alpha)}\int_0^\infty e^{-\theta\left(\frac{\lambda+\alpha}{\lambda}\right)}\theta^{k+\alpha-1}d\theta\end{aligned}$$

$$= \frac{\alpha^\alpha \Gamma(k+\alpha)}{\lambda^\alpha k! \Gamma(\alpha)} \left(\frac{\lambda}{\lambda+\alpha}\right)^{k+\alpha}$$
$$= \frac{\Gamma(\alpha+k)}{\Gamma(\alpha)\Gamma(k+1)} \left(\frac{\alpha}{\lambda+\alpha}\right)^{\alpha} \left(\frac{\lambda}{\lambda+\alpha}\right)^{k} \tag{2.74}$$

Thus, we obtain a negative binomial model of the Negbin II-variety (see equation (2.42)). In order to derive the Negbin I (or Negbin$_k$) model as a Poisson-gamma mixture, the underlying gamma distribution would need to be re-parameterized in a suitable manner.

This type of mixture distribution has an interesting econometric interpretation as the random variation around λ can be thought of as representing unobserved heterogeneity that is likely to arise in most practical situations. For instance, we could have written $\theta = \lambda u$ where λ is deterministic and u is a multiplicative error term. This formulation is equivalent if $u \sim \text{gamma}(\alpha, \alpha)$ and, therefore, $\text{E}(u) = 1$ and $\text{Var}(u) = \alpha^{-1}$.

Whether the gamma distribution is a good model for heterogeneity of the Poisson parameter is a different question. Other mixing distributions are presented later. The great advantage of the gamma distribution is that the integral over the joint distribution can be solved analytically, leading to a mixture distribution (i.e., the negative binomial distribution) in closed form.

2.5.2 Compound Distributions

Compound distributions are of the form

$$Z = \sum_{i=1}^{N} X_i \tag{2.75}$$

where N and $X_i, i = 1, \ldots, N$ are independent random variables. Compound distributions are sometimes also referred to as "stopped-sum distribution" (Santos Silva and Windmeijer, 2001). We say that the distribution of N is generalized by X.

The appeal of this framework is twofold. First, the derivation of the distribution of Z is relatively simple as long as certain independence assumptions are made. Second, and of equal importance, the structure of a compound distribution can be interpreted in terms of an underlying data generating process that has many applications of interest.

Example 1 Let N be the number of families moving from one country or region to another during a given time period, and let X be family size. Z gives then the number of individuals moving.

Example 2 Let N be the number of total events and X be a binary variable that takes the value "1" if the event is reported and "0" otherwise. Z gives then the number of reported events. Similarly, N could be the number of job offers and X a binary indicator that takes the value "1" if the offer is

accepted and the value "0" otherwise. Z represents the number of accepted offers (=job changes).

Example 3 Let N be the number of spells of illness and X be the number of visits to a doctor in a given spell. Z gives then the total number of visits to a doctor.

Example 4 Let X be a degenerate random variable such that $P(X = 1) = 1$. Then $Z = N$.

The concept of compounding is not restricted to cases where N is a count random variable although this is of main interest here. An interesting class of models can be obtained if N is a binary random variable, i.e., Bernoulli distributed, and X is a (truncated) count variable. In one case, without truncation, such a compound distribution is formally equal to a zero-inflated count data model (see Chap. 2.4.3); in the other case, if the generalizing distribution is a truncated-at-zero count data model, we obtain a hurdle model (Santos Silva and Windmeijer, 2001).

The main mathematical tool for studying the properties of compound distributions is again the probability generating function. The following theorem by Feller (1968) gives the key result:

Let $\{X_k\}$ be a sequence of mutually independent random variables with common distribution $P(X_k = i) = f_i$ and probability generating function $\mathcal{P}^{(X)}(s) = \sum_{i=0}^{\infty} f_i s^i$. Consider the sum $Z_N = X_1 + X_2 + \ldots + X_N$, where the number of components N is a random variable independent of X with $P(N = n) = g_n$ and $\mathcal{P}^{(N)}(s) = \sum_{n=0}^{\infty} g_n s^n$. Then

$$\mathcal{P}^{(Z_N)} = \mathcal{P}^{(N)}[\mathcal{P}^{(X)}(s)] \; . \tag{2.76}$$

For a proof, note that the distribution of Z_N can be obtained by the rule for conditional probabilities:

$$h_j = P(Z_N = j) = \sum_{n=0}^{\infty} P[(X_1 + \ldots + X_n) = j]P(N = n) \tag{2.77}$$

For given N, the rule for convolution applies:

$$\mathrm{E}(s^{Z_n}|N = n) = [\mathrm{E}(s^X)]^n \; , \tag{2.78}$$

and the probability generating function of Z_N is equal to the *marginal* expectation

$$\mathcal{P}^{(Z_N)} = \mathrm{E}(s^{Z_N}) = \sum_{n=0}^{\infty} g_n[\mathrm{E}(s^X)]^n = \mathcal{P}^{(N)}[\mathcal{P}^{(X)}(s)] \tag{2.79}$$

which was to be shown. Of particular interest is the case where N has a Poisson distribution function. Z_N then has a compound Poisson distribution with probability generating function

$$\mathcal{P}(s) = e^{-\lambda + \lambda \mathcal{P}^{(X)}(s)} \tag{2.80}$$

Different results follow:

1. Let X_i be identically and independently Bernoulli distributed with $B(1,p)$ and let N have a Poisson distribution function with $Po(\lambda)$. Then Z_N is Poisson distributed with parameter λp. This result follows, since the probability generating function of the Bernoulli distribution has the form $\mathcal{P}(s) = q + ps$. The probability generating function of Z_N is then given by:

$$\begin{aligned}\mathcal{P}(s) &= e^{-\lambda+\lambda(q+ps)} \\ &= e^{-\lambda p+\lambda ps}\end{aligned} \tag{2.81}$$

   Z_N is Poisson-distributed with expectation $\mathrm{E}(Z_N) = \lambda p$. The process of generalizing a Poisson distribution by a Bernoulli distribution is also called "binomial thinning".
2. Let X have a *logarithmic distribution* with parameter θ and let N have a Poisson distribution with parameter λ. Then Z_N is negative binomial distributed with parameters $-\lambda/\log(1-\theta)$ and $\theta/(1-\theta)$. Following the same line of reasoning as before, start with the probability generating function of the logarithmic distribution, which is given by

$$\mathcal{P}(s) = -\alpha[\log(1-\theta s)] \tag{2.82}$$

   with $\alpha = -[\log(1-\theta)]^{-1}$.
   Thus, the probability generating function of Z_N is obtained as

$$\begin{aligned}\mathcal{P}(s) &= \exp[-\lambda - \lambda\alpha\log(1-\theta s)] \\ &= \exp(-\lambda)(1-\theta s)^{-\alpha\lambda} \\ &= [\exp(1/\alpha)]^{-\alpha\lambda}(1-\theta s)^{-\alpha\lambda} \\ &= \left[\frac{1-\theta s}{1-\theta}\right]^{-\alpha\lambda} \\ &= \left[1 - \frac{\theta}{1-\theta}(1-s)\right]^{-\alpha\lambda}\end{aligned}$$

   This is the probability generating function of a negative binomial distribution with parameters $\alpha\lambda = -\lambda/\log(1-\theta)$ and $\theta/(1-\theta)$ (see (2.37)).

Finally, we note that mixing and compounding are related concepts. For instance, consider a mixture distribution of the form (2.67), where

1. F^j are distribution functions of the $j-$fold convolutions of X, and
2. the a'_js are given by the probability function of N.

This is exactly the form of compounding described above. Alternatively, consider a parametric mixture distribution (2.69). Let X be a random variable with probability generating function $[\mathcal{P}^{(X)}(s)]^{\theta}$, where θ is a parameter. Suppose θ itself is a random variable with generating function $\mathcal{P}^{(\theta)}(s)$. Then, the probability generating function of the mixture distribution is given by:

$$\sum [\mathcal{P}^{(X)}(s)]^{\theta} p(\theta) = \mathcal{P}^{(\theta)}[\mathcal{P}^{(X)}(s)] \tag{2.83}$$

which is the generating function of a compound distribution.

2.5.3 Birth Process Generalizations

A pure birth process is defined by the transition probability (see Chap. 2.2.4)

$$P\{N(0, t + \Delta) = k + 1 | N(0, t) = k\} = \lambda\Delta + o(\Delta)$$

The transition probabilities can be used to construct the marginal distribution of the count data $N(T)$. This requires the solution of differential equations of the sort encountered in the context of the Poisson process (Chap. 2.2.3), which is always possible, if not analytically then numerically.

The main property of a pure birth process is that the probability of an event depends on the number of events that have occurred up to that moment, and not on *when* they occurred. The nature of the dependence can be kept very general. In fact, it can be shown that for any count data distribution there exists a sequence $\lambda_0, \lambda_1, \lambda_2, \ldots$ such that the count distribution is generated by the specified birth process (Faddy, 1997).

Thus, rather than specifying a parametric probability function directly, one can instead model the function $\lambda_k = f(k; \theta)$ parametrically and derive the corresponding probability function. A class of particular interest is generated by the function (Faddy, 1997)

$$\lambda_k = a(b + k)^c$$

where $a, b > 0$. This formulation nests the Poisson distribution (for $c = 0$) and the negative binomial distribution (for $c = 1$), and it allows for general types of overdispersion (for $c > 0$) and underdispersion (for $c < 0$) that are not linked to any particular existing parametric distribution.

The use of this model in regression analysis requires an expression for the mean. While the exact mean can in general not be computed analytically, Faddy (1997) derives the following approximation

$$\tilde{\mathrm{E}}(X) = b \left\{ \left[1 + \frac{a(1-c)}{b^{1-c}} \right]^{\frac{1}{1-c}} - 1 \right\} \tag{2.84}$$

The approximation is exact for $c = 0$ and for $c \to 1$. In order to set the (approximate) mean equal to a given value, say μ, one has to parametrize a accordingly, i.e, solve (2.84) for a :

$$a = \frac{(\mu + b)^{1-c} - b^{1-c}}{1 - c}$$

This generalized count data distribution has two more parameters than the Poisson distribution. As in the standard Poisson model, μ can be expressed in terms of covariates, and the parameters of the model can be estimated by maximum likelihood.

2.5.4 Katz Family of Distributions

Distributions for non-negative integers can be uniquely represented by their *recursive probability ratios*

$$\frac{P(X=k)}{P(X=k-1)} = \frac{p_k}{p_{k-1}} = f(k,\theta) \quad k = 1, 2, \ldots \tag{2.85}$$

where θ is a vector of parameters. (2.85) is a first order difference equation of the form $p_k = f(k,\theta)p_{k-1}$.

Different recursive probability systems have been developed. The Katz family of distributions (Johnson and Kotz, 1969, p. 37) is the most prominent among them. The family provides a particularly useful tool for econometric modeling since it constitutes a generalization nesting several distributions for non-negative integers, while maintaining a parsimonious parameterization (two parameters). It is defined by the recursive probabilities

$$\frac{p_k}{p_{k-1}} = \frac{\omega + \gamma(k-1)}{k} \qquad k = 1, 2, \ldots \tag{2.86}$$

Since the right-hand-side has to be positive for all possible values of k, the following restrictions hold: a) $\omega > 0$, and b) $k \leq \omega/\gamma$ for $\gamma < 0$. The Poisson distribution is obtained for $\gamma = 0$, the negative binomial distribution for $0 < \gamma < 1$ and the binomial distribution for $\gamma < 0$ when $-\omega/\gamma$ is an integer. Tab. 2.2 compares the parameterizations:

Table 2.2. Sub-Models of the Katz System

Poisson	$\omega = \lambda\,,\ \gamma = 0$
Negative Binomial	$\omega = \alpha\left(\frac{1}{1+\theta}\right)\,,\ \gamma = \frac{1}{1+\theta}$
Geometric	$\omega = \gamma = \frac{1}{1+\theta}$
Binomial	$\omega = \frac{np}{1-p}\,,\ \gamma = -\frac{p}{1-p},\ y \leq n$

The mean of the Katz family of distributions can be calculated as follows: re-writing (2.86) as

$$kp_k = [\omega + \gamma(k-1)]p_{k-1},\ k = 1, 2, \ldots$$

and taking sums on both sides, one obtains (the derivation in Johnson and Kotz 1969, p. 37, contains an error: The summation in their formula (32) is with respect to j, not r):

$$\begin{aligned}\mathrm{E}(X) &= \sum_{k=1}^{\infty} k p_k \\ &= \omega \sum_{k=1}^{\infty} p_{k-1} + \gamma \sum_{k=1}^{\infty} (k-1) p_{k-1} \\ &= \omega + \gamma \mathrm{E}(X)\end{aligned}$$

and hence

$$\mathrm{E}(X) = \omega/(1-\gamma) \tag{2.87}$$

The second noncentral moment is

$$\begin{aligned}\mathrm{E}(X^2) &= \textstyle\sum_{k=1}^{\infty} [\omega + \omega(k-1) + \gamma(k-1)^2 + \gamma(k-1)] p_{k-1} \\ &= \omega + \gamma \mathrm{E}(X^2) + (\omega + \gamma)\mathrm{E}(X) \\ &= \omega(1+\omega)/(1-\gamma)^2\end{aligned} \tag{2.88}$$

and the variance is given by

$$\begin{aligned}\mathrm{Var}(X) &= \mathrm{E}(X^2) - [\mathrm{E}(X)]^2 \\ &= \omega/(1-\gamma)^2\end{aligned} \tag{2.89}$$

The Katz system of distributions exhibits *equidispersion* for $\gamma = 0$, *overdispersion* for $0 < \gamma < 1$ and *underdispersion* for $\gamma < 0$.

Alternative recursive probability systems have been formulated. For instance, Yousry and Srivastava (1987) include a third parameter to obtain a *hyper-negative binomial distribution*. It is based on the recursive structure

$$\frac{p_{k+1}}{p_k} = \frac{\omega + \gamma k}{k + \psi} \qquad k = 0, 1, \ldots \tag{2.90}$$

For $\psi = 1$, the model corresponds to the Katz family. For $\gamma = 0$, the model reduces to the *hyper-Poisson distribution* mentioned in Johnson and Kotz (1969, p. 43).

Kulasekera and Tonkyn (1992) formulate a probability distribution for strictly positive integers with

$$\frac{p_{k+1}}{p_k} = \left(\frac{k+1}{k}\right)^{\alpha} q \qquad k = 1, 2, \ldots \tag{2.91}$$

where $q \in (0, 1)$ and $\alpha \in \mathbb{R}$. It nests the shifted negative binomial, the logarithmic and the discrete Pareto distribution.

2.5.5 Additive Log-Differenced Probability Models

Gourieroux and Monfort (1990) define the *additive-log-differenced probability class* (A.L.D.P.) which applies to situations, where the function f determining the ratio of recursive probabilities in (2.85) is separable:

$$f(x, \psi) = g(x)h(\psi) \tag{2.92}$$

It is a generalization since ψ can be estimated *without* specifying the functional form of $g(x)$.

The A.L.D.P. class of probability models includes all linear-exponential family defined over the non-negative integers. These families have probability functions of the form

$$f(x; \psi) = c(x, \phi) \exp\{(x\psi - b(\psi))/\phi\}$$

which implies that the ratio of recursive probabilities is given by

$$\frac{f(x; \psi)}{f(x-1; \psi)} = c(x, \phi)/c(x-1, \phi) \exp(\psi) = g(x, \phi)h(\psi)$$

Example 1.
The Poisson distribution is a linear exponential family with $c(x, \phi) = 1/x!$, $b(\psi) = \exp(\psi)$, and $\psi = \log(\lambda)$, where λ is equal to the expected value (mean parameterization). The ratio of recursive probabilities λ/x is separable into two functions $g(x) = 1/x$ and $h(\lambda) = \lambda$.

Example 2.
The geometric distribution is a linear exponential family with $c(x, \phi) = 1$, $b(\psi) = -\log(1 - \exp(\psi))$, and $\psi = \log(\lambda/(1 + \lambda))$, again in mean parameterization with $E(X) = \lambda$. The ratio of recursive probabilities $\lambda/(1 + \lambda)$ is separable into two functions $g(x) = 1$ and $h(\lambda) = \lambda/(1 + \lambda)$.

The A.L.D.P. defines a class of probability distributions whose recursive probability ratio is separable. It is convenient to specify this distribution class in terms of log-ratios. From (2.85) and (2.92) it follows immediately that

$$\log f(x) - \log f(x-1) = \tilde{g}(x) + \tilde{h}(\psi) \tag{2.93}$$

where $\tilde{g} = \log(g)$ and $\tilde{h} = \log(h)$.

2.5.6 Linear Exponential Families

Finally, we present in this section some results on the general distribution class of *linear exponential families* (LEF). The Poisson distribution is part of the linear exponential families, as are, among others, the normal distribution and the binomial distribution. The interest in this class of distributions, in the context of count data modeling, is not related to any avenue for more general distributions that it may suggested (which, in fact, it does not). Rather, the importance of LEFs stems mainly from certain results on robust (or *semi-parametric*) estimation that apply to such models. The density or probability functions of LEF distributions are of the form (see, for instance, McCullagh and Nelder (1989) and Gourieroux, Monfort and Trognon (1984a))

$$f(x;\psi) = c(x,\phi)\exp\{(x\psi - b(\psi))/\phi\}\ . \tag{2.94}$$

ψ is called the "natural" parameter and ϕ the "dispersion" parameter. The functional form (2.94) shows why the name "linear exponential" is appropriate: f is log-linear in its parameter ψ. As a consequence, the derivative of the logarithmic density with respect to ψ has a very simple form

$$\frac{\partial \ln f(x;\psi)}{\partial \psi} = \frac{x - b'(\psi)}{\phi} \tag{2.95}$$

This derivative plays an important role in estimation, as it is the "score" or "gradient" of the log-likelihood function.

It is a well known result in statistics that the expected gradient is zero. This result follows because differentiating the identity

$$\frac{\partial}{\partial \psi}\left(\int f(x;\psi)dx = 1\right)$$

yields (under suitable regularity conditions that ensure that integration and differentiation can be interchanged)

$$\begin{aligned}\int \frac{\partial f(x;\psi)}{\partial \psi}\,dx &= \int \frac{\partial \ln f(x;\psi)}{\partial \psi} f(x;\psi)\,dx \\ &= \mathrm{E}\left(\frac{\partial \ln f(x;\psi)}{\partial \psi}\right) \\ &= 0\end{aligned} \tag{2.96}$$

Applying this result to the right hand side expression of (2.95), we find that

$$\mathrm{E}(X) = b'(\psi) \tag{2.97}$$

The fact that the score of a linear exponential family is the difference between the random variable and its mean constitutes the reason why consistency of the maximum likelihood estimator requires only that the mean $b'(\psi)$ of the LEF is correctly specified (and that ψ is identified). Under independent sampling, the empirical gradient converges in probability to its expected value of zero. But this means that $\hat{\psi}$ converges in probability to the value where $\mathrm{E}(X) = b'(\psi)$. Note that this result is unaffected by the presence of a dispersion parameter ϕ. More details are given in the next chapter, when generalized linear models and robust Poisson regression are introduced.

We conclude by deriving the variance of a LEF distribution. To do this, one can use the results that the expected second derivative of a logarithmic density function is equal to the variance of its first derivative. For a LEF density,

$$\mathrm{E}\left(\frac{\partial^2 \ln f(x;\psi)}{\partial \psi^2}\right) = \frac{-b''(\psi)}{\phi} \tag{2.98}$$

Further, the variance of the first derivative is given by

Table 2.3. Linear Exponential Families

Distribution	$c(x,\phi)$	$b(\psi)$	ψ
Poisson	$1/x!$	$\exp(\psi)$	$\log(\lambda)$
Negative Binomial[a]	$\frac{\Gamma(\alpha+x)}{\Gamma(\alpha)\Gamma(x+1)}$	$-\alpha\log(1-e^{\psi})$	$\log\left(\frac{\theta}{1+\theta}\right)$
Geometric ($\alpha=1$)	1	$-\log(1-e^{\psi})$	$\log\left(\frac{\theta}{1+\theta}\right)$
Binomial (n given)	$\binom{n}{x}$	$n\log(1+e^{\psi})$	$\log\left(\frac{p}{1-p}\right)$
Normal	$\frac{\exp(-\frac{x^2}{2\phi^2})}{\sqrt{2\pi\phi^2}}$	$\psi^2/2$	$\psi(\sigma=\phi)$

	$\mathrm{E}(X)$	$V(\mu)$
Poisson	λ	μ
Negative Binomial[a]	$\alpha\theta$	$\mu+\alpha^{-1}\mu^2$
Geometric ($\alpha=1$)	θ	$\mu+\mu^2$
Binomial (n given)	np	$\mu(1-\mu)$
Normal	ψ	1

[a] for given α

$$\mathrm{E}\left(\frac{\partial \ln f(x;\psi)}{\partial\psi}\right)^2 = \mathrm{E}\left(\frac{x-b'(\psi)}{\phi}\right)^2 = \frac{\mathrm{Var}(X)}{\phi^2} \tag{2.99}$$

It follows that

$$\mathrm{Var}(X) = \phi b''(\psi) \tag{2.100}$$

The variance of X is the product of two components. One, $b''(\psi)$, is called the variance function. It can be written as a function of the mean $\mathrm{E}(X)=\mu$, since from (2.97) it holds that $\psi = (b')^{-1}(\mu)$. The second componet is a scaling parameter ϕ. For instance, the normal distribution assumes a scaling parameter $\phi=\sigma^2$ and a constant variance function $V(\mu)=1$.

Tab. 2.3 gives the characteristics of some common univariate distributions contained in the linear exponential families. Further members are the gamma and the inverse Gaussian distributions.

2.5.7 Summary

This section has introduced different types of generalizations. These generalizations had as a common point of departure the Poisson distribution with its

restrictive assumptions and structure. More flexible probability models have been developed along two different routes.

The first approach formulated compound and mixture distributions the development of which often was motivated by a reconsideration of the data generating process. For instance, a compound Poisson distribution can describe the number of migrants between two geographical areas if the number of *families* moving is Poisson distributed and the number of *persons* in each family follows a binomial distribution. A Poisson mixture distribution may be appropriate if the Poisson parameter λ is measured with error.

The second approach directly formulated a more richly parameterized, and therefore more general, distribution model. An example is the Katz system of distributions. If the interest of the researcher rests less in the best possible fit to observed data but rather in the robustness of parameter estimates, distributions within the class of linear exponential families have desirable robustness properties.

2.6 Distributions for Over- and Underdispersion

2.6.1 Generalized Event Count Model

The generalized event count model as presented in King (1989b) and extended in Winkelmann and Zimmermann (1991) is essentially a re-parametrization of the Katz family discussed in Chap. 2.5.4, in terms of a mean λ and a variance function $g(\lambda)$ that allows for a simple test for overdispersion or underdispersion. Moreover, since the model is parameterized in terms of the mean, it can be readily extended to a regression context, for instance by expressing λ as a function of regressors.

Recall that the Katz family is defined by a recursive formula for the probabilities $f(y)$ (where the notation differs slightly from the previous chapter):

$$\frac{f(y+1)}{f(y)} = \frac{\theta + \gamma y}{1+y} \text{ for } y = 0, 1, 2, \ldots \text{ and } \theta + \gamma y \geq 0\,. \tag{2.101}$$

Using recursive substitution, (2.101) can be rewritten as

$$f(y|\theta, \gamma) = f(0) \prod_{j=1}^{y_i} \left[\frac{\theta + \gamma(j-1)}{j} \right], y_i = 1, 2, \ldots \tag{2.102}$$

where $f(0)$ is determined by the fact that the probabilities have to sum to one. Mean and variance are then given by

$$\mathrm{E}(y) = \frac{\theta}{(1-\gamma)}\,,\ \mathrm{Var}(y) = \frac{\theta}{(1-\gamma)^2} \tag{2.103}$$

It is easily seen that this family produces equidispersion for $\gamma = 0$, overdispersion for $0 < \gamma < 1$, and underdispersion for $\gamma < 0$. The following parameterization has been suggested.

$$\gamma = \frac{(\sigma^2-1)\lambda^k}{(\sigma^2-1)\lambda^k+1}\ ,\ \theta = \frac{\lambda}{(\sigma^2-1)\lambda^k+1} \tag{2.104}$$

With this parameterization,

$$\mathrm{E}(y) = \lambda$$

and

$$\mathrm{Var}(y) = \lambda + (\sigma^2-1)\lambda^{k+1}$$

Special case of interest are $\sigma^2 = 1$, the Poisson case. For $\sigma^2 < 1$, we obtain underdispersion, for $\sigma^2 > 1$, overdispersion. The variance function can be linear ($k = 0$), or quadratic ($k = 1$), restrictions that can be tested. The complete probability function is given by:

$$f_{\mathrm{geck}}(y|\cdot) = f(0|\lambda,\sigma^2,k) \times \begin{cases} \prod_{j=1}^{y}\left[\frac{\lambda+(\sigma^2-1)\lambda^k(j-1)}{[(\sigma^2-1)\lambda^k+1]\,j}\right] & \text{for } y = 1,2,\ldots \\ 1 & \text{for } y = 0 \end{cases} \tag{2.105}$$

where

$$f(0|\lambda,\sigma^2,k) = \begin{cases} (1+(\sigma^2-1)\,\lambda^k)^{\nu} & \text{for } \sigma^2 \geq 1 \\ (1+(\sigma^2-1)\,\lambda^k)^{\nu} D^{-1} & \text{for } 0<\sigma^2<1\,, \\ & \lambda^k \leq 1/(1-\sigma^2) \\ & \text{and } y \leq \mathrm{int}^*(\nu) \\ 0 & \text{otherwise} \end{cases}$$

$$\nu = \lambda^{1-k}/(1-\sigma^2)\,,$$

$$D = \textstyle\sum_{m=0}^{\mathrm{int}^*(\nu)} f_{\mathrm{bn}}(m|\lambda,\sigma^2,k)\,,$$

$$\text{and } \mathrm{int}^*(\mathrm{y}) = \begin{cases} \mathrm{int(y)}+1 & \text{for } \mathrm{int(y)}<\mathrm{y} \\ \mathrm{y} & \text{for } \mathrm{int(y)}=\mathrm{y} \end{cases}$$

The limit of $f(0|\lambda,\sigma^2,k)$ for $\sigma^2 \to 1$ is $e^{-\lambda}$ and the GEC_k converges to the Poisson model.

2.6.2 Generalized Poisson Distribution

A detailed presentation of the *generalized Poisson distribution* can be found in a monograph by Consul (1989). Further references are Consul and Famoye (1992), Famoye (1993), and Wang and Famoye (1997). The latter references explicitly introduce exogenous variables and thus a generalized Poisson regression model. Santos Silva (1997b) extended the model to truncated data.

The generalized Poisson distribution allows for both over- and underdispersion and nests the Poisson regression model as a special case. This is achieved by introducing one additional parameter θ. The probability distribution function can be written as (Consul 1989, p. 4)

$$f(y) = \begin{cases} \dfrac{\theta(\theta + y\gamma)^{y-1} e^{-\theta - y\gamma}}{y!}, & y = 0, 1, 2, \ldots \\ 0 \qquad \text{for } y > m\,, \text{ when } \gamma < 0 \end{cases} \tag{2.106}$$

where $\theta > 0$, $\max[-1, -\theta/m] < \gamma \leq 1$ and $m(\geq 4)$ is the largest positive integer for which $\theta + m\gamma > 0$ when γ is negative.

The generalized Poisson distribution nests the Poisson distribution for $\gamma = 0$. Mean and variance are given by $\mathrm{E}(y) = \theta(1-\gamma)^{-1}$ and $\mathrm{Var}(y) = \theta(1-\gamma)^{-3}$, respectively. Thus, the generalized Poisson distribution displays overdispersion for $0 < \gamma < 1$, equidispersion for $\gamma = 0$ and underdispersion for $\max[-1, -\theta/m] < \gamma \leq 0$. Therefore, the parameter space is restricted in case of underdispersion.

To obtain the model in mean parameterization, let

$$\theta = \frac{\lambda}{1 + a\lambda}$$

$$\gamma = \frac{a\lambda}{1 + a\lambda}$$

Now, the probability function can be written as

$$f(y) = \left(\frac{\lambda}{1 + a\lambda}\right)^{y} \frac{(1 + ay)^{y-1}}{y!} \exp\left(-\frac{\lambda(1 + ay)}{1 + a\lambda}\right) \tag{2.107}$$

and the mean and variance of y are given by

$$\mathrm{E}(y) = \lambda$$

$$\mathrm{Var}(y) = \lambda(1 + a\lambda)^2$$

When $a = 0$, (2.107) reduces to the standard Poisson distribution. a acts like a dispersion parameter, with underdispersion for $a < 0$ and overdispersion for $a > 0$.

2.6.3 Poisson Polynomial Distribution

Cameron and Johansson (1997) discuss a class of parametric models for count data, using a squared polynomial expansion around a Poisson distribution, based on work by Gallant and Nychka (1987). Guo and Trivedi (2002) derive a corresponding polynomial expansion of the negative binomial distribution. See also Romeu and Vera-Hernandez (2005).

If $f(y; \lambda)$ is a Poisson or a negative binomial distribution, a new probability distribution may be obtained by letting

$$g_p(y;\lambda,a) = f(y;\lambda)\frac{[h_p(y;a)]^2}{\eta_p(\lambda,a)} \qquad y = 0,1,2,\ldots \tag{2.108}$$

where

$$h_p(y;a) = \sum_{k=0}^{p} a_k y^k \ ,$$

$p = 1,2,\ldots$ describes the order of the polynomial, $\eta_p(\lambda,a)$ is a normalizing constant that ensures that the density $g_p(y;\lambda,a)$ sums to one, and squaring the polynomial ensures that the probabilities are non-negative. The parameters of the extended model are then estimated by maximum likelihood.

Using this method, one can approximate the unknown true probability function arbitrarily closely, by increasing the polynomial order. In fact, it can be shown that the mean and other aspects of the unknown probability are estimated consistently provided that the length of the series increases with sample size. This holds regardless of the baseline distribution, which gives the method a non-parametric flavor, although the interest usually centers on estimation of parameters. Thus, although this method uses maximum likelihood, consistent estimation does not require that the baseline distribution is correctly specified.

Cameron and Johansson (1997) show that the normalizing constant is of the general form

$$\eta_p(\lambda,a) = \sum_{k=0}^{p}\sum_{l=0}^{p} a_k a_l m_{k+l}$$

where $m_r = m_r(\lambda)$ denotes the r-th non-central moment of the baseline density $f(y;\lambda)$.

For example, a Poisson distribution expanded by a squared polynomial of order one has probability function (the constant of the polynomial is normalized to unity)

$$g_1(y;\lambda,a_1) = \frac{e^{-\lambda}\lambda^y}{y!}\frac{(1+a_1 y)^2}{\eta_1(\lambda,a_1)} \tag{2.109}$$

where

$$\eta_1(\lambda,a_1) = m_0 + 2a_1 m_1 + a_1^2 m_2 = 1 + 2a_1\lambda + a_1^2(\lambda+\lambda^2)$$

The distribution has therefore one additional parameter, a_1, more than the baseline Poisson probability function. This additional parameter helps relaxing the equidispersion property of the Poisson distribution, as $g_1(y;\lambda,a_1)$ can be overdispersed, equidispersed (for $a_1 = 0$), or underdispersed. The implied flexibility appears to be a useful feature, making the Poisson polynomial model a potentially useful candidate distribution when one is not sure *a-priori*, whether the population model is over- or underdispersed. It does not require any restrictions on the parameter space, or an upper limit on the counts, as

did for instance the generalized event count model and the generalized Poisson distribution discussed earlier.

On the downside, while the model is well suited to fit a model to actual data, it is less straightforward to extend it to a regression framework. The reason is that the mean is a function of both λ and α, which prohibits a simple, intuitive interpretation of parameters as marginal mean effects. Testing for over- or underdisperion is also complicated, as it cannot be pinned to a simple point hypothesis on α_1. Whether the distribution is overdispersed or underdispersed depends on both α and λ. We also note a further slightly awkward aspect of the model, which is that the set of possible outcomes is not necessarily equal to the natural numbers plus zero: whenever $y = -1/a_1$ that particular y has probability zero. For all of these reasons, and possibly also due to computational complexities associated with multiple local maxima of the log likelihood function, this approach has not been used extensively in the follow-up literature.

2.6.4 Double Poisson Distribution

The double Poisson distribution has been proposed by Efron (1986). The distribution has two parameters, λ and θ, and its probability function is:

$$f(y, \lambda, \theta) = K(\lambda, \theta)\sqrt{\theta}\exp(-\lambda\theta)\exp(-y)\frac{y^y}{y!}\left(\frac{e\lambda}{y}\right)^{\theta y} \tag{2.110}$$

where

$$K(\lambda, \theta) \approx 1 + \frac{1-\theta}{12\lambda\theta}\left(1 + \frac{1}{\lambda\theta}\right)$$

and $\lambda, \theta > 0$. For $\theta = 1$, the double Poisson distribution collapses to the simple Poisson distribution. The advantage of the double Poisson distribution is that it introduces one additional parameter, θ and the variance and mean are no longer necessarily equal. Efron (1986) shows that

$$\mathrm{E}(y) \approx \lambda$$

$$\mathrm{Var}(y) \approx \lambda/\theta$$

Hence, the double Poisson distribution allows for overdispersion when $\theta < 1$, and for underdisersion when $\theta > 1$. A disadvantage of this distribution is that these results are not exact, as the normalizing constant is not available in closed form. As a consequence, as for the Poisson polynomial distribution, the first moment is not available in closed form as well.

2.6.5 Summary

Table 1.1 of the previous Chapter showed some empirical count data distributions. While most of them were overdispersed at the marginal level, there was

one exception, the number of children, where underdispersion was observed. In the absence of a-priori information whether a count is overdispersed or underdispersed, it is clearly desirable to have access to a class of models that can accommodate both over- and underdispersion at the same time, without imposing any a-priori restrictions. Such probability distributions were discussed in this section.

It turned out that finding such a distribution with good statistical properties is not straightforward. Each of the distributions here had an aspect that might be considered a down-side. For two distributions, the range of the admissible observations depended on the parameter value. For two distributions, the exact means and variances were not available in closed forms. One distribution for over- and underdispersion that avoids these difficulties, the gamma-count distribution (Winkelmann, 1995), is discussed in Chap. 2.7.3. Its derivation requires results from duration analysis and renewal processes, which are presented next.

2.7 Duration Analysis and Count Data

When looking at a sequence of events, most econometricians are more familiar with the concept of *waiting times* (or, in technical applications: *failure* times) and *duration models* than with the concept of event counts. See Allison (1984) for an excellent introduction to duration models. Lancaster (1990) provides a more advanced treatment. The count and the duration view are just two different representations of the same underlying stochastic process. Understanding the interlinkages between the two provides a deeper understanding of the assumptions and specification issues involved in count data analysis.

The key insight is that the distributions of cumulative waiting times uniquely determine the distribution of counts, and they are, in turn, uniquely determined by the distribution of counts. This isomorphism can be exploited to derive new count data distributions, as in Winkelmann (1995), Lee (1996), and Bradlow et al. (2006), and to improve our understanding of the properties of count data models in general. In particular, a new interpretation can be given to the presence of over- and underdispersion based on the duration properties of the underlying process.

The fundamental relationship between counts and durations was introduced in Chap. 2.2.6. It is repeated here for convenience. Let $N(T)$ denote the total number of events that have occurred between 0 and T, and let ϑ_k denote the arrival time of the k-th event. Then by definition

$$N(T) < k \text{ if and only if } \vartheta_k > T \tag{2.111}$$

and

$$P(N(T) < k) = P(\vartheta_k > T) = 1 - F_k(T)\,, \tag{2.112}$$

where $F_k(T)$ is the cumulative density function of ϑ_k. Further,

$$\begin{aligned} P(N(T) = k) &= P(N(T) < k+1) - P(N(T) < k) \\ &= P(\vartheta_{k+1} > T) - P(\vartheta_k > T) \\ &= F_k(T) - F_{k+1}(T) \end{aligned} \tag{2.113}$$

where $F_k(T)$ is the cumulative distribution function of ϑ_k and it is understood that $F_0(T) = 1$. Equation (2.113) provides the fundamental relation between the distribution of waiting times and the distribution of counts. The probability distribution of $N(T)$ can be obtained explicitly for all k from knowing the distribution of ϑ_k. Similarly, we can solve 2.113 for $F_k(T)$ to obtain

$$F_1(T) = 1 - P(N(T) = 0)$$

$$F_2(T) = F_1(T) - P(N(T) = 1) = 1 - P(N(T) = 0) - P(N(T) = 1)$$

and, in general

$$F_k(T) = 1 - \sum_{j=0}^{k-1} P(N(T) = j)$$

We can now study, for example, the arrival time distributions implied by the Poisson assumption. Since

$$P(N(T) = k) = \frac{e^{-\lambda T}(\lambda T)^k}{k!}$$

where λ is the expected number of events per time unit, the arrival time of the first event is then

$$F_1(T) = 1 - e^{-\lambda T}$$

Thus, ϑ_1 has an exponential distribution with parameter λ and density function

$$f(\vartheta_1) = \lambda e^{-\lambda T}$$

Similarly, the arrival time of the second event, again assuming that the counts follow a Poisson distribution, is given by

$$F_2(T) = 1 - e^{-\lambda T} - \lambda T e^{-\lambda T}$$

with density

$$f(\vartheta_2) = \lambda e^{-\lambda T} - (\lambda e^{-\lambda T} - \lambda^2 T e^{-\lambda T}) = \lambda^2 T e^{-\lambda T}$$

This generalizes further, and we obtain

$$F_k(T) = 1 - \sum_{j=0}^{k-1} \frac{e^{-\lambda T}(\lambda T)^j}{j!}$$

with first derivative

$$f_k(T) = \frac{\lambda^k T^{k-1} e^{-\lambda T}(\lambda T)^j}{(k-1)!} \tag{2.114}$$

This is the probability function of the *Erlang distribution*. To recapitulate, we have the result that if $N(T)$ is Poisson distributed, then we know that the distribution of the arrival time of the k-th event must be the Erlang distribution.

To further study the stochastic implications of the Poisson assumption, it is instructive to focus on the interarrival times τ_i, rather than the arrival times themselves. Per definition, the interarrival time τ_i is the time elapsed between the occurrence of the $(i-1)$'th and the i'th event. Formally,

$$\tau_k = \vartheta_k - \vartheta_{k-1}$$

and

$$\vartheta_k = \sum_{i=1}^{k} \tau_i \tag{2.115}$$

Clearly, $\vartheta_1 = \tau_1$, and we thus know that if the counts are Poisson distributed, τ_1 must be exponential distributed. What we would like to know is what kind of sequence of τ_i's would lead to the Poisson distribution, and then of course also, what kind of count distribution will arise for non-exponentially distributed interarrival times. We start, in the next Chapter, with some definitions, characterizations and properties of distributions for interarrival times. We then introduce, in Chapter 2.7.2, the concept of a renewal process. This framework provides a tractable approach to link distributions for interarrival times and counts under a variety of distributional assumptions, exponential being one of them, albeit under the restrictive set-up of independent and identical interarrival distributions.

2.7.1 Distributions for Interarrival Times

Interarrival times are non-negative continuous random variables, denoted as τ. $f(t)$ is the *density function* of the interarrival time, $F(t) = P(\tau < t)$ is the *distribution function*, and $\bar{F}(t) = 1 - F(t)$ is the *survivor function*. An important entity for the analysis of durations, used to capture duration dependence, is the *hazard rate* $\lambda(t)$ which gives the instantaneous exit probability conditional on survival. Formally,

$$\lambda(t) = \lim_{dt \to 0} \frac{P(t \leq \tau < t + dt | \tau \geq t)}{dt} \tag{2.116}$$

Using Bayes rule

$$\begin{aligned} P(t \leq \tau < t + dt | \tau \geq t) &= \frac{P(t \leq \tau < t + dt, \tau \geq t)}{P(\tau \geq t)} \\ &= \frac{P(t \leq \tau < t + dt)}{P(\tau \geq t)} \end{aligned}$$

Expressing the probabilities through cumulative density functions, dividing by dt and taking limits, we obtain

$$\lambda(t) = \frac{f(t)}{1 - F(t)} = -\frac{d}{dt} \log \bar{F}(t) \tag{2.117}$$

The hazard function captures the underlying time dependence of the process. A decreasing hazard function implies that the spell is less likely to end the longer it lasts. This situation is referred to as *negative duration dependence*. An increasing hazard function implies that the spell is the more likely to end the longer it lasts. This situation is referred to as *positive duration dependence*. Clearly, positive duration dependence reduces the probability of long spells, while negative duration dependence does the opposite. No duration dependence corresponds to the case of a constant hazard. The hazard function of an interarrival time distribution does not need to be monotonic although many parametric distributions that are used in practice have a monotonic hazard function.

In the case of a constant hazard $\lambda(t) = \lambda$, there is a unique underlying distribution which can be obtained directly as the solution to the differential equation

$$-\frac{d}{dt} \log \bar{F}(t) = \lambda \tag{2.118}$$

Hence, $\bar{F}(t) = Ae^{-\lambda t}$, and, using the initial condition $\bar{F}(0) = 1$,

$$\bar{F}(t) = e^{-\lambda t} \tag{2.119}$$

Thus, under the assumption of a constant hazard function we obtain that $F(t) = 1 - e^{-\lambda t}$ and $f(t) = \lambda e^{-\lambda t}$. These are the distribution and the density functions of the exponential distribution, respectively. The hazard (which equals the inverse of the expected value) is constant if and only if the distribution of completed spells is exponential.

A constant hazard function is closely related to the concept of absence of memory within the process: A process is called *memoryless* if the probability that a spell exceeds $s + t$, conditional on having lasted for t, is independent of t.

Formally, this can be written as

$$P\{\tau > s + t | \tau > t\} = P\{\tau > s\} \quad \forall\, s, t \geq 0\ . \tag{2.120}$$

In other words, the past length of a spell is without influence on its future duration. Using Bayes' rule, this condition can be rewritten as

$$\frac{P\{\tau > s + t, \tau > t\}}{P\{\tau > t\}} = P\{\tau > s\} \tag{2.121}$$

or, since $\{\tau > t\}$ is contained in $\{\tau > s + t\}$,

$$P\{\tau > s + t\} = P\{\tau > t\} P\{\tau > s\}\ . \tag{2.122}$$

The survivor function of the exponential distribution is given by $P\{\tau > t\} = e^{-\lambda t}$. Since $e^{-\lambda(t+s)} = e^{-\lambda t} e^{-\lambda s}$, it follows that exponentially distributed

waiting times is memoryless. If the spell lasted until time t, the distribution of the remaining elapsed time until completion is identical to the original lifetime distribution.

Clearly, the assumption of a constant hazard is too restrictive in most applications. Distributions that allow for positive or negative duration dependence are, among others, the gamma and the Weibull distributions. Both distributions are characterized by a monotonic hazard function, either increasing or decreasing, the slope of which depends on the value taken by a specific parameter.

So far the discussion has focused on the distribution of interarrival times τ_i. The distributions of the arrival times ϑ_k are obtained by convolution, as in (2.115). For example, if τ_1 and τ_2 are independently exponential distributed with same parameter λ, and if $f_2(T)$ denotes the density function of the arrival time of the second event, then we can write

$$f_2(T) = \int_{t=0}^{T} \lambda e^{-\lambda t} \lambda e^{-\lambda(T-t} dt = \lambda^2 T e^{-\lambda T}$$

But this is exactly the density function of the Erlang distribution for $k = 2$, and the argument indeed generalizes to $k > 2$, as shown in Chapter 2.2.6. But as $N(T)$ is Poisson distributed if an only of ϑ_k is Erlang distributed, we know that independent and identical exponentially distributed interarrival times lead to a number of events that is Poisson distributed.

This is a special case of a *renewal process*, i.e., a stochastic process that excludes inter-spell dependence and assumes i.i.d. interarrival times. Renewal processes may, however, display duration dependence and in the following section, results from renewal theory are used to provide interesting insights in the relationship between duration dependence and the distribution of counts.

2.7.2 Renewal Processes

Useful references on renewal processes are Barlow and Proschan (1965), Cox (1962), Feller (1971), and Lancaster (1990). Consider a stochastic process that is defined by a sequence of spells τ_i, where the end of one spell immediately leads to the start of a new spell. If $\{\tau_1, \tau_2, \ldots\}$ are independently and identically distributed variables, all with density function $f(\tau)$, the process is called a *renewal process*. Let $N(T)$ denote the number of renewals in $(0, T)$, i.e., the number of events before T, a count variable. Its probability function in terms of the cumulative densities of arrival times ϑ_k was given in (2.113). But $\vartheta_k = \sum_{i=1}^{k} \tau_i$. Given the assumption of independent renewals, the distribution of this k-fold convolution can be derived using the calculus of Laplace transforms. In general, the Laplace transform of ϑ, denoted as is k-th power of the Laplace transform of τ:

$$\mathcal{L}_{\vartheta_k}(s) = [\mathcal{L}_\tau(s)]^k \tag{2.123}$$

(See Feller, 1971). In some cases, we can start with a well-known parametric distribution function for τ and obtain through convolution the Laplace transform of another distribution that is again of recognizable form. An extreme case would be a family of distributions that is "closed under convolution", i.e. for which distributions of sums belong to the same family of distributions as the components. Example for such cases are the Normal distribution, the Poisson distribution, the binomial distribution and the gamma distribution. Of these, only the gamma distribution is a useful candidate as a model for τ, as we require a distribution for a non-negative continuous variable representing time.

In other cases, however, it will be very hard if not impossible to derive the distribution of ϑ_k using (2.123). It turns out, however, that a useful limiting result relating the hazard function of τ_i and the distribution of counts can be obtained even without fully specifying the distribution of τ.

Denote the mean and the variance of the waiting time distribution by $\mathrm{E}(\tau) = \mu$ and $\mathrm{Var}(\tau) = \sigma^2$, and the coefficient of variation by $v = \sigma/\mu$. Assume that the (unknown) distribution of τ has a monotonic hazard function, such that $\mathrm{d}\lambda(t)/\mathrm{d}t$ is either positive, zero, or negative for all values of t. Thus, we allow for the three cases of positive, negative, or no duration dependence.

Barlow and Proschan (1965, p. 33) have shown the type of duration dependence puts bounds the coefficient of variation of the distribution of τ. In particular,

$$\left.\begin{matrix} < \\ = \\ > \end{matrix}\right\} 0 \Longrightarrow v \left.\begin{matrix} > \\ = \\ < \end{matrix}\right\} 1$$

A second important result due to Cox (1962, p.40) is that if $\{\tau_i\}$ is a sequence of independent, positive, identically distributed interarrival times with mean μ and variance σ^2, then $N(t)$, the number of renewals (or counts) is asymptotically normal distributed with mean t/μ and variance $\sigma^2 t/\mu^3$:

$$N(t) \overset{asy}{\sim} \text{normal}\left(\frac{t}{\mu}, \frac{\sigma^2 t}{\mu^3}\right) \tag{2.124}$$

as $t \to \infty$.

As a consequence, we know that the ratio of variance to mean of the limiting distribution is given by

$$\frac{\mathrm{Var}(N(t))}{\mathrm{E}(N(t))} = \frac{\sigma^2 t}{\mu^3}\frac{\mu}{t} = \frac{\sigma^2}{\mu^2} = v^2 \tag{2.125}$$

Thus, we can link the duration dependence of the underlying interarrival time distributions to the dispersion of the counts, via the coefficient of variation of the distribution of τ. The variance mean ratio is greater (less) than 1 if and only if the coefficient of variation of the waiting times $v = \sigma/\mu$ is greater (less) than 1. For positive duration dependence $v < 1$ and the count

distribution is underdispersed. For negative duration dependence $v > 1$ and the count distribution is overdispersed.

The exponential distribution has coefficient of variation $v = 1$, leading to equidispersion. This result is exact, whereas (2.125) is only a limiting result.

2.7.3 Gamma Count Distribution

The renewal framework can also be used to derive an exact distribution for event counts, based on gamma distributed renewals (Winkelmann, 1995). This is possible since the gamma distribution is, as mentioned earlier, closed under convolution. It is also interesting since the gamma distribution has a monotonic hazard function that is either increasing, constant, or decreasing. Under the gamma assumption, the density of τ is given by

$$f(\tau;\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\tau^{\alpha-1}e^{-\beta\tau}\,, \qquad \alpha,\,\beta \in I\!R^+ \tag{2.126}$$

with Laplace transform

$$\mathcal{L}_\tau(s) = (1+s/\beta)^{-\alpha} \tag{2.127}$$

The waiting time has mean $\mathrm{E}(\tau) = \alpha/\beta$ and variance $\mathrm{Var}(\tau) = \alpha/\beta^2$. The hazard function $\lambda(\tau)$ obeys the equation

$$\frac{1}{\lambda(\tau)} = \int_0^\infty e^{-\beta u}\left(1+\frac{u}{\tau}\right)^{\alpha-1}\mathrm{d}u \tag{2.128}$$

The gamma distribution admits no closed form expression for the tail probabilities and thus no simple formula for the hazard function. However, from (2.128), it follows that $\lambda(\tau)$ is (monotonically) increasing for $\alpha > 1$, decreasing for $\alpha < 1$, and constant (and equal to β) for $\alpha = 1$.

Now, consider the distribution of ϑ_k, the arrival time of the k-th event. Applying (2.123), we find that

$$\mathcal{L}_{\vartheta_k}(s) = (1+s/\beta)^{-\alpha k} \tag{2.129}$$

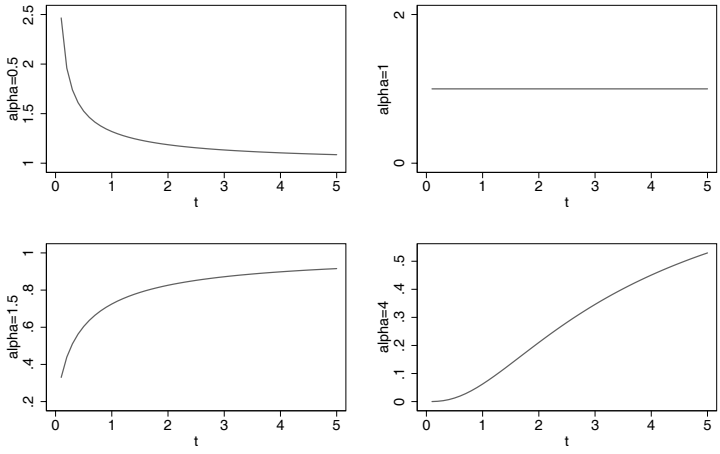
which is the Laplace transform of a gamma distribution with parameters β and αk. We can thus directly write the density function of ϑ_k as

$$f_k(\vartheta;\alpha,\beta) = \frac{\beta^{k\alpha}}{\Gamma(k\alpha)}\vartheta^{k\alpha-1}e^{-\beta\vartheta} \tag{2.130}$$

To derive the new count data distribution, we have to evaluate the cumulative distribution function

$$\begin{aligned} F_k(T) &= \int_0^T \frac{\beta^{k\alpha}}{\Gamma(k\alpha)}\vartheta^{k\alpha-1}e^{-\beta\vartheta}d\vartheta \\ &= \frac{1}{\Gamma(k\alpha)}\int_0^{\beta T} u^{k\alpha-1}e^{-u}\mathrm{d}u \end{aligned} \tag{2.131}$$

Fig. 2.3. Hazard Rates for Gamma Distribution ($\beta = 1$)



where the second equality uses the change of variable to $u = \alpha\vartheta$. The right-hand side is an incomplete gamma integral that will be denoted as $G(\alpha k, \beta T)$. For non-integer α, no closed-form expression is available for $G(\alpha k, \beta T)$ (and thus for $P(N = k)$). Numerical evaluations of the integral can be based on asymptotic expansions (See Abramowitz and Stegun, 1964, and Bowman and Shenton, 1988).

The number of event occurrences during the time interval $(0, T)$ has then the two-parameter distribution function

$$P(N(T) = k) = G(\alpha k, \beta T) - G(\alpha k + \alpha, \beta T) \qquad k = 0, 1, 2, \ldots \qquad (2.132)$$

where it is understood that $F_0(T) = G(0, \beta T) = 1$.

For $\alpha = 1$, $f(\tau)$ is the exponential density and (2.132) simplifies to the Poisson distribution. For $0 < \alpha < 1$, the gamma count distribution is based on interarrival times with negative duration dependence. For $\alpha > 1$, the duration dependence is positive. Fig. 2.4 and 2.5 compare the probability functions of the gamma count distribution with a Poisson distribution of identical mean ($\mathrm{E}(N) = 2$) for two values of α. Depending on the value of α, the Gamma count model is more concentrated ($\alpha = 1.5$) or more dispersed ($\alpha = 0.5$) than the reference distribution.

The gamma count distribution is one of the few distributions that nests the Poisson distribution through a parametric restriction and allows for both over- and underdispersion. In contrast to the other two such distributions discussed in this book, the generalized event count model and the generalized Poisson distribution that have been covered earlier in Chapters 2.6.1 and 2.6.2, respectively, it has the additional advantage that it does not impose

Fig. 2.4. Probability Functions for Gamma Count and Poisson Distributions; $\alpha = 0.5$ (Overdispersion)

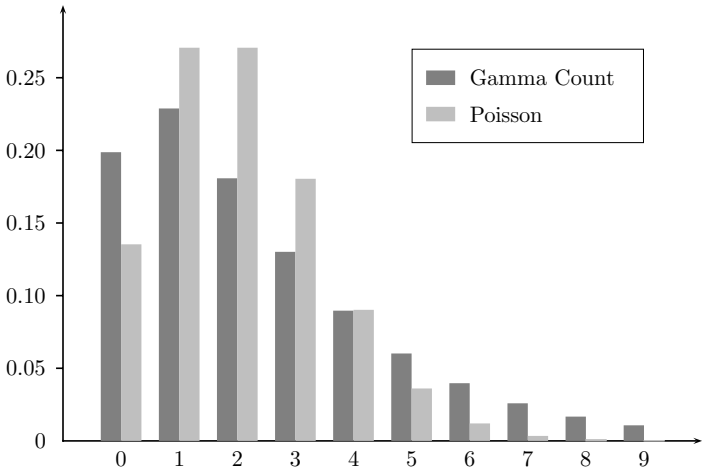
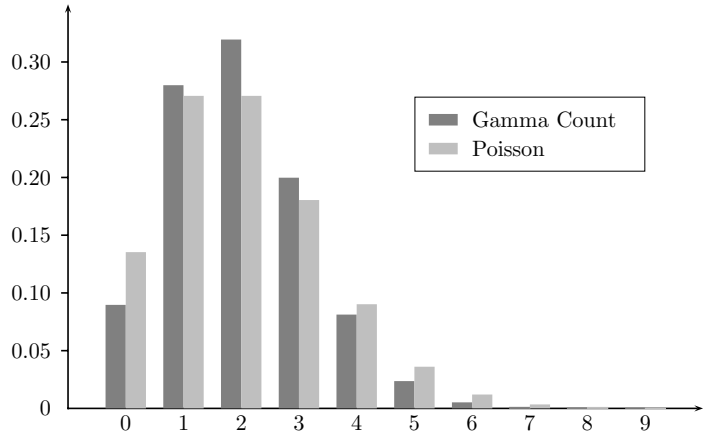


Fig. 2.5. Probability Functions for Gamma Count and Poisson Distributions; $\alpha = 1.5$ (Underdispersion)



any restrictions on the range of the outcome variables. Both generalized event count model and generalized Poisson distribution impose an upper limit for the outcome variable in the case of underdispersion. Seen from the point of view of a renewal process, there is no reason for such an asymmetry, and this approach allows for a unified treatment of over- and underdispersion. Clearly, this feature is also a potentially great advantage over the negative binomial model, where only overdispersion is possible and the Poisson model lies at the boundary of the parameter space.

There is a small catch, though, which is that the expected value is not available in closed form. Rather, it must be computed as

$$\begin{aligned} \mathrm{E}[N(T)] &= \sum_{k=1}^{\infty} kP(N(T)=k) \\ &= \sum_{k=1}^{\infty} k[G(\alpha k, \beta T) - G(\alpha k + \alpha, \beta T)] \\ &= \sum_{i=1}^{\infty} G(\alpha k, \beta T) \end{aligned} \tag{2.133}$$

We will discuss in a later Chapter, how this kind of model can be transformed into a regression model. The variance is given by

$$\mathrm{Var}[N(T)] = \sum_{k=1}^{\infty} k^2[G(\alpha k, \beta T) - G(\alpha k + \alpha, \beta T)] - \mathrm{E}[N(T)]^2 \tag{2.134}$$

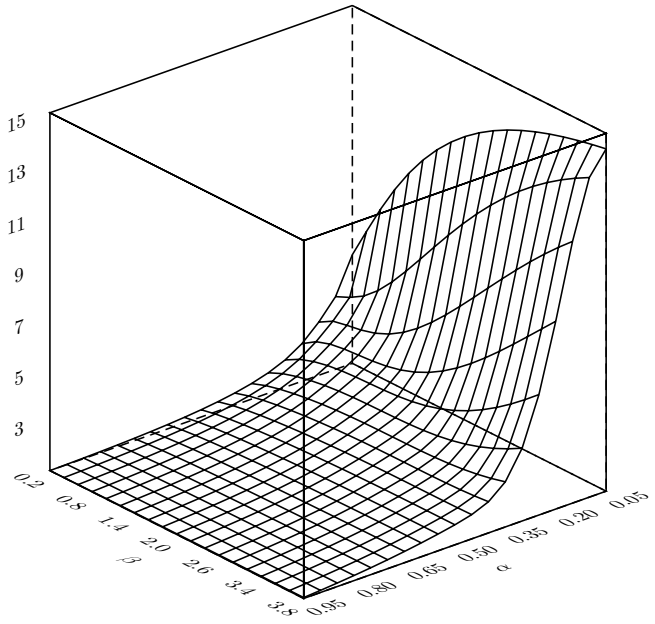
As previously indicated, the gamma count distribution function (2.132) displays overdispersion for $0 < \alpha < 1$ and underdispersion for $\alpha > 1$. Fig. 2.6 and Fig. 2.7 show the variance mean ratio for various values of α and β.

Recall from above that the underlying waiting times have a decreasing (increasing) hazard for $0 < \alpha < 1$ ($\alpha > 1$). Thus, as in the limiting case of a renewal process considered above, negative duration dependence leads to overdispersion, positive duration dependence to underdispersion. The intermediate case of no duration dependence, i.e., exponentially distributed waiting times, leads to the Poisson distribution with equal mean and variance.

2.7.4 Duration Mixture Models

The phenomena of a positive or negative relationship between duration and hazard in the aggregate does not need to reflect 'true' duration dependence but can also be due to a selection process: to take the example of negative duration dependence, individuals with duration of spells above average might have a (constant) hazard below average. Failure to account for this heterogeneity, for example by splitting up the population into sub-populations, results in spurious negative duration dependence. The problem of identifying true duration dependence was discussed in detail by Heckman and Singer (1984).

Fig. 2.6. Variance to Mean Ratio for Gamma Count Distribution; $0 < \alpha < 1$



It is closely related to the problem of distinguishing between occurrence dependence and unobserved heterogeneity in count data.

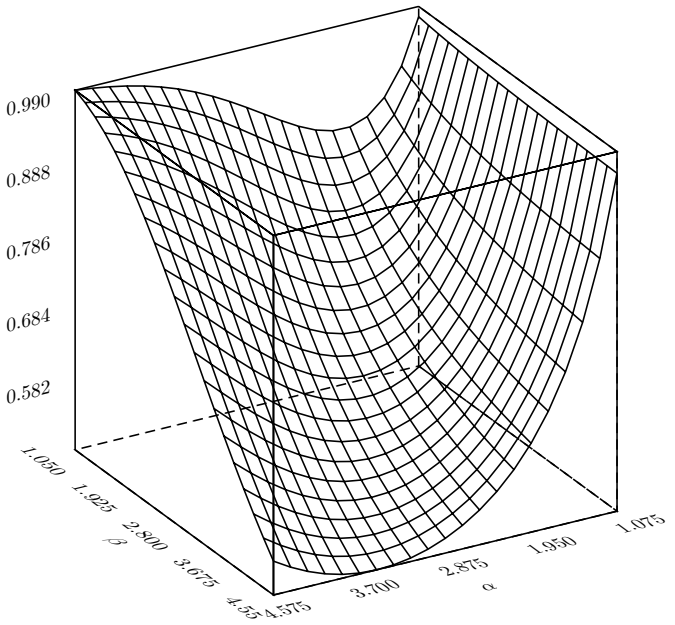
If one suspects the presence of unobserved heterogeneity, one possible solution is to assume that the heterogeneity follows a specific parametric distribution. This was already discussed in Chap. 2.5.1 where it was shown that if the Poisson parameter λ is a random variable with a gamma distribution, the number of events occurring in a given interval has a negative binomial distribution Negbin(α, θ). A corresponding result exists in the duration domain:

Assume that the Poisson parameter λ is a random variable with a gamma distribution. Then the waiting time for the first occurrence has an exponential-gamma mixture distribution and the hazard rate $\lambda(t) = \alpha/(\beta + t)$ is a decreasing function of time.

This result holds, since the probability function of the count is obtained in the usual way via integration, where we keep explicitly track of varying lengths of the time interval:

$$P(X = k; t) = \int_0^\infty \frac{(\lambda t)^k e^{-\lambda t}}{k!} \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\lambda\beta} d\lambda$$

Fig. 2.7. Variance to Mean Ratio for Gamma Count Distribution; $\alpha > 1$



$$= t^k \frac{\Gamma(\alpha+k)}{\Gamma(\alpha)\Gamma(k+1)} \left(\frac{\beta}{\beta+t}\right)^{\alpha} \left(\frac{1}{\beta+t}\right)^{k}$$

But $P(X = 0; t)$ is the probability that no event has occurred before t, i.e., the survivor function at t:

$$\bar{F}(t) = \left(\frac{\beta}{\beta+t}\right)^{\alpha} \tag{2.135}$$

with density function

$$f(t) = -\frac{d}{dt}\bar{F}(t) = \alpha\beta^{\alpha}\left(\frac{1}{\beta+t}\right)^{\alpha+1} \tag{2.136}$$

and hazard rate

$$\begin{aligned}\lambda(t) &= -\frac{d}{dt}\log\bar{F}(t) \\ &= \frac{\alpha}{\beta+t} \\ &= \frac{\lambda}{1+\lambda/\alpha t}\end{aligned} \tag{2.137}$$

where the last line follows from letting $\lambda = \alpha/\beta$.

Thus, if $\alpha \to \infty$ and $1/\beta \to 0$ such that $\alpha/\beta = \lambda$ (i.e., the gamma mixture distribution has a mean of λ and a variance approaching zero) the hazard function collapses to the hazard function of the exponential distribution.

Incidentally, the density (2.136) can also be obtained directly by mixing the exponential density of the arrival time of the first event with a gamma distribution (see also Lancaster, 1990, Chap. 4):

$$\begin{aligned} f(t) &= \int_0^\infty \lambda e^{-\lambda t} \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\lambda\beta} d\lambda \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty \lambda^\alpha \lambda e^{-\lambda(t+\beta)} d\lambda \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha+1)}{(t+\beta)^{\alpha+1}} \end{aligned} \tag{2.138}$$

With heterogeneity, the sample hazard is no longer constant but instead decreasing with time. As indicated above, this model describes a situation where each individual has a constant hazard that randomly varies between individuals according to a gamma law. The gamma disturbance captures unobserved heterogeneity. In terms of counts it leads to the negative binomial distribution with overdispersion. In terms of waiting times it leads to a decreasing overall hazard since the mean hazard among *survivors* at t is a decreasing function of t. A selection effect is taking place: individuals with larger hazard are likely to exit earlier, and the group of survivors is increasingly composed of individuals with relatively small λ's. The parameters α and β have opposite effects. An increase in α increases the hazard proportionally for all lengths of duration. The negative effect of β is reduced with increased duration.

3

Poisson Regression

3.1 Specification

3.1.1 Introduction

The Poisson regression model is the benchmark model for count data in much the same way as the normal linear model is the benchmark for real-valued continuous data. Early references in econometrics include Gilbert (1982), Hausman, Hall and Griliches (1984), and Cameron and Trivedi (1986). The Poisson model is simple, and it is robust. If the only interest of the analysis lies in estimating the parameters of a log-linear mean function, there is hardly any reason (except for efficiency) to ever contemplate anything other than the Poisson regression model. In fact, its applicability extends well beyond the traditional domain of count data. The Poisson regression model can be used for any constant elasticity mean function, whether the dependent variable is a count or continuous, and there are good reasons why it should be preferred over the more common log transformation of the dependent variable.

And yet, there are instances where the Poisson regression model is unsuited. Essentially, the Poisson model is always overly restrictive when it comes to estimating features of the population other than the mean, such as the variance or the probability of single outcomes. The simplicity of the Poisson regression model, an asset when modeling the mean, turns then into a liability, and more elaborate models are needed.

3.1.2 Assumptions of the Poisson Regression Model

The basic Poisson regression model relates the probability function of a dependent variable y_i (also referred to as regressand, endogenous, or dependent variable) to a vector of independent variables x_i (also referred to as regressors, exogenous, or independent variable). Let k be the number of regressors (including, usually, a constant). x_i is then a column vector of dimension $(k \times 1)$. Finally, n is the number of observations in the sample.

The standard univariate Poisson regression model makes the following three assumptions:

Assumption 1

$$f(y|\lambda) = \frac{e^{-\lambda}\lambda^{y}}{y!} \qquad y = 0, 1, 2, \ldots$$

where $f(y|\lambda)$ is the conditional probability function of y given λ, and it must hold that $\lambda > 0$.

Assumption 2

$$\lambda = \exp(x'\beta)$$

where β is a $(k \times 1)$ vector of parameters, and x is a $(k \times 1)$ vector of regressors, including a constant.

Assumption 3

observation pairs $(y_i, x_i), i = 1, \ldots, n$ are independently distributed.

Discussion

Assumptions 1 and **2** can be combined to obtain the following conditional probability function:

$$f(y|x) = \frac{\exp(-\exp(x'\beta))\exp(yx'\beta)}{y!} \qquad y = 0, 1, 2, \ldots \tag{3.1}$$

The Poisson distribution has only one parameter that simultaneously determines conditional mean and variance. Therefore, the Poisson regression model as defined by the assumptions above implies an exponential (or log-linear) mean function,

$$\mathrm{E}(y|x) = \lambda = \exp(x'\beta) \tag{3.2}$$

and an exponential conditional variance function

$$\mathrm{Var}(y|x) = \lambda = \exp(x'\beta) \tag{3.3}$$

The fact that conditional mean and conditional variance are equal in the Poisson regression model is a particular feature – equidispersion – that will be subject to further discussion.

The probabilistic assumptions underlying the Poisson distribution have been discussed in the previous chapter. In a nutshell, events are assumed to occur truly randomly over time. In the context of the regression model, explanatory variables influence the dependent variable (the number of event counts in a time interval) through the intensity (or instantaneous occurrence

rate) of the process. The heterogeneity of the latter is modeled as a deterministic function of the explanatory variables. This implies that, unlike for the normal linear regression model, the randomness of the Poisson model is intrinsic and not due to an additive stochastic error representing additional heterogeneity.

If the underlying stochastic process does not display the required randomness, or if there is not even a meaningful underlying stochastic process to think of, the Poisson regression model may remain a valid approximation to the true data generating process as well as a useful descriptive tool.

In conjunction with **Assumptions 1** and **2**, **Assumption 3** allows for a straightforward application of the method of maximum likelihood to estimate the parameters of the model. Maximum likelihood estimation is discussed below.

3.1.3 Ordinary Least Squares and Other Alternatives

The advantages and disadvantages of the Poisson regression model are best contemplated by addressing the practitioner's question "When and why should the Poisson regression model be used?". A natural first answer would seem to be that the dependent variable should be a count. But this condition is neither necessary nor sufficient.

It is not necessary, because the Poisson regression model has been shown to be useful for non-count dependent variables as well. One example is the exponential regression model with right censoring, which arises in continuous time duration modeling, and which can be estimated by Poisson regression. Another, more important example, is the estimation of any constant elasticity model by Poisson regression. This application is discussed at greater length in Chapter 3.3.5.

Obviously, the fact that the dependent variable is a count is not sufficient either. Firstly, there are alternative count data models that take the nature of the dependent variable into account and that may be superior to the Poisson model. Often, such generalized models will allow for a richer set of inferences, in particular with respect to the probability of single outcomes (such as "zero") and with it on the underlying structural data generating process. Possible specifications of alternative count data models and the selection of the right model are important topics covered in later chapters.

Secondly, it is not obvious, why one cannot ignore the special nature of the dependent variable altogether and just apply standard regression models such as the normal linear model

$$y = x'\beta + e \qquad e|x \sim \mathrm{N}(0, \sigma^2) \tag{3.4}$$

Several objections against such an approach can be brought forward. (3.4) ignores the discrete nature of the dependent variable. Under the normal linear model, the probability of any particular outcome is zero. Thus, no inferences on single outcomes are possible.

In addition, model (3.4) allows for negative outcomes whereas counts are non-negative. And relatedly, the model violates **Assumption 2** that the mean function is log-linear. Thus, (3.4) will give an inconsistent estimator of β if the true data generating process follows the Poisson regression model. Finally, it ignores the heteroskedasticity inherent in count data (see equation (3.3)). The only vindication of this approach arises if counts are very large. The Poisson distribution, for example, can be approximated by a normal distribution, and the approximation is usually deemed satisfactory for $\lambda > 20$.

Log-Linear Model

These concerns can in part be addressed by conventional methods. Start with the mean function. We could specify

$$\log y = x'\beta + u \qquad u \sim \mathrm{N}(0, \sigma^2) \tag{3.5}$$

where "log" denotes the natural logarithm. In this model,

$$y = \exp(x'\beta + u) \tag{3.6}$$

has a log-normal distribution with conditional expectation

$$\mathrm{E}(y|x) = \exp(x'\beta + 1/2\sigma^2)$$

similar to the Poisson regression model (up to a scale factor $\exp(1/2\sigma^2)$), and the values of y are restricted to the non-negative real line. As long as the model has an overall constant, we can redefine $\tilde{\beta}_0 = \beta_0 - 1/2\sigma^2$ and the two models have essentially the same mean function.

The log-normal distribution implies a different variance function, though. In particular, it holds that

$$\mathrm{Var}(y|x) = \phi[\mathrm{E}(y|x)]^2$$

where $\phi = e^{\sigma^2} - 1$. In general, estimated standard errors of the log-normal and Poisson models won't be comparable, and heteroskedasticity consistent standard errors should be computed.

The two fundamental problems with the log-normal approach are that "zero" counts are inadmissible, as the logarithm is defined only for positive outcomes, and that a "re-transformation" problem arises: if the conditional variance of y is not quadratic in the conditional expectation, the log-linear model provides an inconsistent estimator of the semi-elasticities of interest (see Chapter 3.3.5).

Ad-hoc solutions to the zero problem have been proposed, such as dropping all zero outcomes, or adding a constant c, such as 0.1 and 0.5, to each count (see King, 1988). In this case, the model is written as

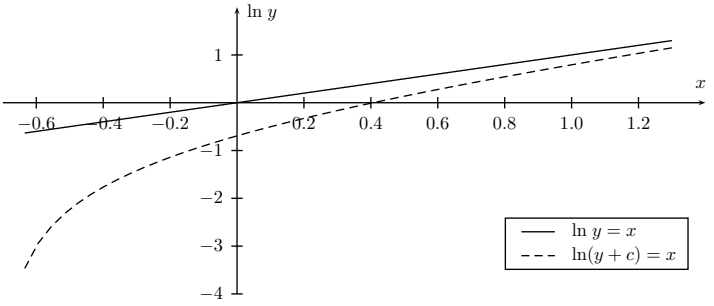
$$\log(y + c) = x'\beta + u$$

or, equivalently.

$$\log y = \log(\exp(x'\beta + u) - c)$$

which shows that $\log y$ is now non-linear in x. Estimation of such a model in general will introduce bias, as illustrated in Fig. 3.1. In this figure, models with and without adjustment ($c = 0.5$) are drawn for $u = 0$ in the $\log y/x$ space (assuming that there is only one regressor and $\beta_0 = 0$ and $\beta_1 = 1$). As is apparent from Fig. 3.1, the slope parameter of the adjusted models exceeds the true slope of unity the more, the closer the value of x comes to its logical lower bound $\log(c)$.

Fig. 3.1. Bias in the Log-Linear Model When a Constant is Added in Order to Deal With Zero Counts



King (1988) reports results from a Monte Carlo analysis where the adjusted log-linear model is applied to artificial data from a Poisson regression model. He finds substantial bias for the parameter estimates when the log-linear model is used instead of the Poisson regression model. The bias does not disappear with increasing sample size. The log-linear model tends to overestimate the slope parameters when positive, and to underestimate the slope parameters when negative, i.e., the parameters are biased away from zero. By introducing bias and ignoring the discrete nature of the data, this model is quite unsatisfactory and its use cannot be recommended. Similarly, of course, dropping all zeros is not a good idea either, as it will lead to endogenous sample selection problems similar to those known from the linear model (Heckman, 1979).

Non-Linear Least Squares

Part of the problem arises because we have considered a model with multiplicative error $\varepsilon = \exp(u)$ (See equation (3.6)). Consider the alternative model

$$y = \exp(x'\beta) + v \qquad v \overset{\text{i.i.d}}{\sim} \mathrm{N}(0, \sigma^2) \tag{3.7}$$

This model has the same mean function as the Poisson regression model. As the logarithm of y needs not to be taken, the problem with zero counts disappears. Indeed, from an estimation point of view, model (3.7), if estimated by maximum likelihood (which is the same as non-linear least squares in this case), gives a consistent estimator of β if the true world is Poisson (estimation is discussed in greater detail later in the book). It is not efficient, as it ignores the heteroskedasticity inherent in the Poisson regression model. However, the model could be modified such that $\sigma^2 = \exp(x'\beta)$, in which case iteratively weighted non-linear least squares would give the same results as Poisson maximum likelihood. The main problem with this model is that, while offering no advantage in terms of ease of estimation or interpretation of parameters, it fails to take into account the non-negative and integer-valued nature of the dependent variable. The model cannot be used to predict the probability of single outcomes.

Ordered Probit and Logit

Two non-count data models that overcome these shortcomings are the ordered logit and ordered probit models. Both are models for experiments in which outcomes are measured on an ordinal scale. An example is a survey question that solicits the agreement or disagreement with a certain proposition (such as: X is a good teacher) using the responses *strongly disagree / disagree / neutral / agree / strongly agree*. The five possible outcomes can be coded, for instance, as 0,1,2,3, and 4, respectively, although the coding is arbitrary as long as it preserves the ordering.

The models are based on an underlying latent model

$$y^* = x'\beta + \varepsilon$$

with the observation mechanism

$$\begin{aligned} &y = 0 \text{ if } y^* < \alpha_0 \\ &y = 1 \text{ if } \alpha_0 \leq y^* < \alpha_1 \\ &y = 2 \text{ if } \alpha_1 \leq y^* < \alpha_2 \\ &\vdots \end{aligned}$$

where α_j are "threshold values". Depending on the assumptions for ε, the ordered probit ($\varepsilon \sim N(0,1)$) or ordered logit ($\varepsilon \sim$ standard logistic) arises. Given α, β and x, the probability of each of the 5 possible outcomes is determined and α and β can be estimated by maximum likelihood. Clearly, ordinal models can also be used for counts as long as the number of different counts observed in the sample is not too large. The number of threshold parameters that require estimation increases with the observed sample space by one-to-one; for more details on ordered response models, see McKelvey and Zavoina (1975), and Boes and Winkelmann (2006).

Ordered models in general provide a better fit to the data than pure count data models. The threshold parameters give the flexibility to align predicted

and actual frequencies. However, their use for modeling count data has a number of serious deficiencies.

- They are theoretically implausible as a model for counts. They are not based on the concept of an underlying count process.
- Counts are cardinal rather than ordinal. Hence, under the ordinal approach, the sequence "2, 5, 50" is assumed to carry the same information as the sequence "0, 1, 2" which is not the case for count data. Ordinal models disregard this information and cannot be efficient.
- One reason of having parametric models in the first place is the ability of predicting the probability of arbitrary counts. While genuine count data models can do that, ordered models can only predict outcomes that are actually observed in the sample.

In addition, these models in general imply a mean function that is different from the mean function of the standard count model. In the general case, the mean function of ordered models (not the latent model) is highly complex. Consider the simplest case of a binary 0/1 variable. For example, in the binary logit model

$$P(y = 0) = \frac{1}{1 + \exp(x'\beta)}$$

and

$$\mathrm{E}(y|x) = \frac{\exp(x'\beta)}{1 + \exp(x'\beta)}$$

For the Poisson model, instead,

$$P(y = 0) = \exp(-\exp(x'\beta))$$

and

$$\mathrm{E}(y|x) = \exp(x'\beta)$$

The two are fundamentally different mean functions. Of course, this does not imply that the Poisson model is necessarily the superior model, as its mean function may be misspecified as well. However, it suggests that the use of ordered models for count data, and the interpretation of the results, has to proceed with necessary caution. In practice, applications of ordered models to count data are uncommon.

To summarize, the Poisson regression model has many virtues when one wishes to model a count dependent variable. The Poisson model accounts for the discrete and non-negative nature of the data. It attributes positive probability to the outcome "zero". And it allows inferences to be drawn on the probability of a particular outcome. The Poisson regression model naturally accounts for the heteroskedastic and skewed distribution associated with a non-negative random variable. The more the mean of the dependent variable

approaches zero and thus the lower bound of its sample space, the smaller the variance. Finally, the Poisson model has a simple structure and the parameters of the model can be estimated with relative ease.

3.1.4 Interpretation of Parameters

The exponential form of the mean function implies that the necessary increase in $x'\beta = \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik}$ to obtain a one unit increase in $\mathrm{E}(y|x)$ is smaller, the further one moves away from zero. To put it differently, the level change in $x'\beta$ required for a given percentage change in $\mathrm{E}(y|x)$ is kept constant. As a consequence the partial derivative of $\mathrm{E}(y|x)$ with respect to any element of x depends on the value of $x'\beta$:

$$\frac{\partial \mathrm{E}(y|x)}{\partial x_j} = \exp(x'\beta)\beta_j = \mathrm{E}(y|x)\beta_j \qquad j = 1, \ldots, k \tag{3.8}$$

These marginal effects obviously differ between individuals. Sometimes, it may be informative to calculate the marginal effects for a representative individual, such as the sample average of the explanatory variables. Alternatively, one can calculate the expected (or average) marginal effect

$$\mathrm{E}_x \left[\frac{\partial \mathrm{E}(y|x)}{\partial x_j} \right] = \beta_j \mathrm{E}[\exp(x'\beta)]$$

which can be estimated consistently by

$$\widehat{\mathrm{E}}_x \left[\frac{\partial \widehat{\mathrm{E}(y|x)}}{\partial x_j} \right] = \frac{1}{n} \sum_{i=1}^{n} \exp(x'\hat{\beta})\hat{\beta}_j$$

It is more common, and simpler, though, to consider the *relative change* in $\mathrm{E}(y|x)$ associated with a small change in x_j since this is constant and equal to β_j:

$$\frac{\partial \mathrm{E}(y|x)/\mathrm{E}(y|x)}{\partial x_j} = \beta_j \tag{3.9}$$

If x is in logarithmic form, β_j has the interpretation of an elasticity, giving the percentage change in $\mathrm{E}(y|x)$ per percentage change in x_j.

Sometimes, we are interested in assessing the effect of a (discrete) unit change in x_j on the expected value of y. That is, we want to compare the expected value of y for x_j and $x_j + 1$, respectively. In this case, the calculus method gives only an approximation of the relative change. Define $\tilde{x} = (1, x_2, \ldots, x_j + 1, \ldots, x_k)'$. The exact relative change is then

$$\begin{aligned} \frac{\mathrm{E}(y|\tilde{x}'\beta) - \mathrm{E}(y|x'\beta)}{\mathrm{E}(y|x'\beta)} &= \frac{\exp(x'\beta + \beta_j) - \exp(x'\beta)}{\exp(x'\beta)} \\ &= \exp(\beta_j) - 1 \end{aligned}$$

The leading example is that of a dummy variable taking values 0 or 1. Hence, the relative impact of a dummy variable on the expected count is $\exp(\beta_j) - 1$. A linear Taylor series approximation of $\exp(\beta_j) - 1$ around $\beta_j^0 = 0$ yields

$$\begin{aligned}\exp(\beta_j) - 1 &\approx [\exp(\beta_j^0) - 1] + \exp(\beta_j^0)(\beta_j - \beta_j^0)\big|_{\beta_j^0=0} \\ &= \beta_j\end{aligned}$$

Thus, β_j is the first-order approximation to the relative impact of a dummy variable for small β_j, and the linear approximation is the better the smaller β_j.

These results are similar to those encountered in the standard log-linear model. However, there is a conceptual difference that removes a certain ambiguity in the interpretation of the Poisson parameters, an ambiguity that was first noted by Goldberger (1968) for the log-linear model (see also Winkelmann, 2001). There, $\mathrm{E}(\log y|x) = x'\beta$, from which it does not follow that $\mathrm{E}(y|x) = \exp(x'\beta)$. It is only under some additional assumptions that an expression such as $\exp(\beta_j) - 1$ correctly identifies the relative change in $\mathrm{E}(y|x)$ due to a unit change in x_j. The situation in the Poisson regression model is much more straightforward. However, estimation is still an issue. As pointed out by Goldberger (1968) for the log-linear model, estimating $\exp(\beta_j) - 1$ by $\exp(b_j) - 1$, where b_j is the maximum likelihood estimator, though consistent, introduces small sample bias. An improved estimator has been suggested by Goldberger (1968) and Kennedy (1981).

Interactive Effects and Differences in Differences

Interactive terms are used to model complementarities between variables. For instance, let

$$\mathrm{E}(y|x_1, x_2) = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2) \tag{3.10}$$

In this model, $x_1 x_2$ is an interactive term, and β_3 measures its effect on the linear predictor. For instance, for positive β_3, the impact of a given increase in x_1 on $\mathrm{E}(y|x_1, x_2)$ is the larger, the greater is the value of x_2, since

$$\frac{\partial \mathrm{E}(y|x_1, x_2)}{\partial x_1} = \mathrm{E}(y|x_1, x_2)(\beta_1 + \beta_3 x_2) \tag{3.11}$$

In the logic of a multiplicative model, we can then define the *absence* of an interactive effect by requiring that the *relative* change in $\mathrm{E}(y|x_1, x_2)$ associated with a change in x_1 does not depend on x_2 (and vice versa). Dividing equation (3.11) by $\mathrm{E}(y|x_1, x_2)$, and differentiating once more with respect to x_2, we obtain

$$\frac{\partial}{\partial x_2}\left(\frac{\partial \mathrm{E}(y|x_1, x_2)/\partial x_1}{\mathrm{E}(y|x_1, x_2)}\right) = \beta_3 \tag{3.12}$$

If $\beta_3 = 0$, there is no interaction effect, a proposition that can be easily tested. Mullahy (1999) discusses difficulties that arise if interactive terms are defined in terms of absolute (rather than relative) changes, since

$$\frac{\partial^2 \mathrm{E}(y|x_1, x_2)}{\partial x_1 \partial x_2} = \mathrm{E}(y|x_1, x_2) \times [\beta_3(1 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2) + \beta_1 \beta_2] \tag{3.13}$$

Clearly, in this definition, $\beta_3 = 0$ is neither necessary nor sufficient for the absence of an interaction. However, this linear approach is not a natural concept to start with. In the preferred focus on proportional changes, the parameter β_3 has a straightforward interpretation as shown above.

Often x_1 and x_2 are dummy variables. A leading application arises in the treatment effects literature, where x_1 indicates treatment status ($x_1 = 1$ for treatment group and $x_1 = 0$ for control group) and x_2 the observation period ($x_2 = 1$ for post treatment period and $x_2 = 0$ for pre-treatment period). In such a model, the pre-post growth factor in the expected outcomes for the treatment group is

$$\frac{\mathrm{E}(y|x_1 = 1, x_2 = 1)}{\mathrm{E}(y|x_1 = 1, x_2 = 0)} = \exp(\beta_2 + \beta_3)$$

For the control group, the pre-post growth factor in the expected outcomes for the control group is accordingly

$$\frac{\mathrm{E}(y|x_1 = 0, x_2 = 1)}{\mathrm{E}(y|x_1 = 0, x_2 = 0)} = \exp(\beta_2)$$

The identifying assumption for a causal inference is that in the absence of the treatment, the relative change of the expected outcome for the treatment group would have been identical to that actually observed for the control group. The treatment effect TE is then the relative difference between the actual post-treatment outcome of the treatment group, $\mathrm{E}(y|x_1 = 1, x_2 = 1)$ and the counterfactual outcome (the pre-treatment outcome for the treatment group times the growth factor of the control group)

$$\mathrm{E}(y|x_1 = 1, x_2 = 0) \times \frac{\mathrm{E}(y|x_1 = 0, x_2 = 1)}{\mathrm{E}(y|x_1 = 0, x_2 = 0)}$$

which is the same as the ratio of the growth factors of the treatment and control groups minus one:

$$TE = \frac{\dfrac{\mathrm{E}(y|x_1 = 1, x_2 = 1)}{\mathrm{E}(y|x_1 = 1, x_2 = 0)}}{\dfrac{\mathrm{E}(y|x_1 = 0, x_2 = 1)}{\mathrm{E}(y|x_1 = 0, x_2 = 0)}} - 1 = \exp(\beta_3) - 1 \tag{3.14}$$

Thus, the differences-in-differences estimator should be based, in this multiplicative case, on "ratios-in-ratios". In this interpretation, β_3 directly approximates the causal treatment effect (for small β_3).

Marginal Probability Effects

So far, the discussion has focussed exclusively on marginal mean effects, i.e., the question how the mean, or conditional expectation function, varies as any of the explanatory variables changes, ceteris paribus. This focus is natural. It very closely resembles the approach commonly taken in linear regression models for continuous data. However, the focus on conditional expectations is arguably also overly restrictive, as it misses some of the richness inherent in modeling discrete data, and count data in particular. The discreteness implies that statements about single probabilities are meaningful. Such statements may be of substantive interest.

The question then becomes: how does the distribution (or probability function) respond to a small ceteris paribus change in an explanatory variable. The answer to this question is given by the "marginal probability effects". In the exponential Poisson regression model

$$\frac{df(y;\lambda)}{dx} = \frac{df(y;\lambda)}{d\lambda}\frac{d\lambda}{dx} = f(y;\lambda)(y-\lambda)\beta \qquad y = 0, 1, \ldots \tag{3.15}$$

where $f(y;\lambda)$ is the Poisson probability function and $\lambda = \exp(x'\beta)$. We see that the marginal probability effects are quite restrictive. This follows directly from the simple structure of the Poisson model. Observe that

$$\text{sgn}(df(y;\lambda)/dx_j) = -\text{sgn}(\beta_j) \text{ iff } y < \lambda$$

$$\text{sgn}(df(y;\lambda)/dx_j) = \text{sgn}(\beta_j) \text{ iff } y > \lambda$$

Increasing the value of the dependent variable, y, over its support, starting at zero, it must be the case that the marginal probability effects are either initially positive, turning negative after a certain value of the dependent variable; or they are initially negative, turning positive after a certain value. One may refer to this result as a "single-crossing" property of the Poisson model.

Depending on the question one wants to address, models allowing for more flexible marginal probability effects may be desirable. Such models will be discussed later on, one very prominent example being the class of hurdle models (Chap. 6.3). Generally speaking, single-index models (such as the Poisson or negative binomial regression models) will always be restrictive, as the pattern of marginal probability effects is fully determined by the functional form of the underlying probability function - it cannot be modeled flexibly based on covariates and corresponding parameter values.

More flexible models have additional parameters through which the covariates (or linear combinations, or "indices" thereof) can affect the probability function. For example, if a model has two parameters, θ_1 and θ_2, we may let $\theta_1 = g_1(x'\beta_1)$ and $\theta_2 = g_2(x'\beta_2)$. In such models, changes of an explanatory variable may have different effects in different parts of the outcome distribution. Also, by implication, marginal mean effects are then no longer directly linked to marginal effects for higher order moments. For example, the effect

of a variable on the dispersion can be determined independently of its effects on the mean.

3.1.5 Period at Risk

Count data measure the number of times a certain event occurs during a given time interval. The length of this interval is sometimes called "risk-period", or "exposure". In the standard Poisson model, it is assumed that the risk period is the same for all observations. Under this assumption, it can be normalized to unity without loss of generality, and $\exp(x'\beta)$ is the expected value of y per time interval (such as year, month, or week).

However, in other cases, the risk-period varies across observations. For instance, McCullagh and Nelder (1989) analyze the number of reported damage incidents by ship type. Aggregate months of service vary from 45 months for one ship type to 44,882 months for another. Clearly, one would expect the number of incidents to increase with aggregate months of service. In a similar vein, Diggle, Liang and Zeger (1994) use data from a randomized experiment in order to compare the number of epileptic seizures during a 8-week pre-treatment observation period with the number of epileptic seizures during 2-week post-treatment observation period. Finally, Barmby, Nolan and Winkelmann (2001) analyze the number of days absent from work for a sample of workers some of whom are contracted for 4 days of work per week while others are contracted for 5 days of work per week.

Differences in exposure need not be limited to calendar time. For instance, Rose (1990) analyses the determinants of air-traffic incidents. In her case, the different size of operation between the various carriers is expressed by the number of scheduled departures per year (in thousands). Bauer et al. (1998) are interested in the number of workplace accidents, using firm level data for Germany. Again, one would expect that the number of accidents increases with the size of the risk-group, here the number of workers.

The benchmark case for dealing with exposure is to assume proportionality. In the above examples, McCullagh and Nelder (1989) model the expected number of ship damage incidents *per* aggregate month of operation, while Rose (1990) models the expected number of air-traffic incidents *per* 1,000 departures. If we denote the individual level of exposure by t, we can write

$$\frac{\mathrm{E}(y|x)}{t} = \exp(x'\beta) \tag{3.16}$$

or, equivalently,

$$\begin{aligned} \mathrm{E}(y|x) &= t\exp(x'\beta) \\ &= \exp(x'\beta + \log t) \end{aligned} \tag{3.17}$$

Thus, under proportionality, a doubling in exposure time doubles the expected count. In the second line of (3.17), $\log t$ is sometimes referred to as "logarithmic offset".

Alternatively, one might want to give the proportionality assumption free for test. A simple possibility is to include $\log t$ as a regressor without restricting its coefficient to unity:

$$\mathrm{E}(y|x,t) = \exp(x'\beta + \gamma \log t) \tag{3.18}$$

The restriction $H_0 : \gamma = 1$ can then be tested with standard methods. Alternatively, one can reparameterize using $\theta = \gamma - 1$. The mean function then reads

$$\mathrm{E}(y|x,t) = \exp(x'\beta + \theta \log t + \log t) \tag{3.19}$$

Logarithmic time of exposure is included twice, both as offset and as regressor, and the test for proportionality now simplifies to testing $H_0 : \theta = 0$.

Yet another variant of this test exists if time of exposure can take on only two values, t_1 and t_2. This case is presented in Barmby, Nolan and Winkelmann (2001). Define a dummy variable $D = 1$ if $t = t_1$ and $D = 0$ if $t = t_2$. Then, in the regression model

$$\mathrm{E}(y|x,D) = \exp(x'\beta + \delta D) \tag{3.20}$$

the test for proportionality to exposure reduces to $H_0 : \delta = \log(t_1/t_2)$. To establish the equivalence, note that

$$\log t = \log t_2 + (\log t_1 - \log t_2)D$$

Thus, under strict proportionality,

$$\mathrm{E}(y|x,D) = \exp(\log t_2 + x'\beta + (\log t_1 - \log t_2)D)\,,$$

i.e., $\delta = \log t_1 - \log t_2 = \log(t_1/t_2)$ ($\log t_2$ is absorbed into the overall constant).

Endogenous Exposure Time

An interesting alternative class of models arises if the time of exposure is endogenous. A potential source of endogeneity could be that exposure time depends on the occurrence or non- occurrence of events. One plausible scenario is that of a "blockage": there is no exposure, i.e., a zero risk of occurrence or at least of measuring an occurrence, for some given time interval following an occurrence. An example for such a situation occurs in the modeling of fertility: after a birth, no further birth can occur for a period of about 11 months, adding the time of pregnancy and uterine involution.

Feller (1971, p. 372) discusses a blockage time model where events are generated by a Poisson process, and after each event occurrence, a fixed bloackage time ξ sets in. Feller shows that 1) the number of recorded events is not Poisson distributed; and 2) that the number of recorded events is underdispersed. Such a model would be well suited to think about the pregnancy example above. Indeed, it is typically found that fertility data are underdispersed (Winkelmann and Zimmermann, 1994, see also Chap. 9.5 and the empirical fertility distribution in Tab. 1.1)

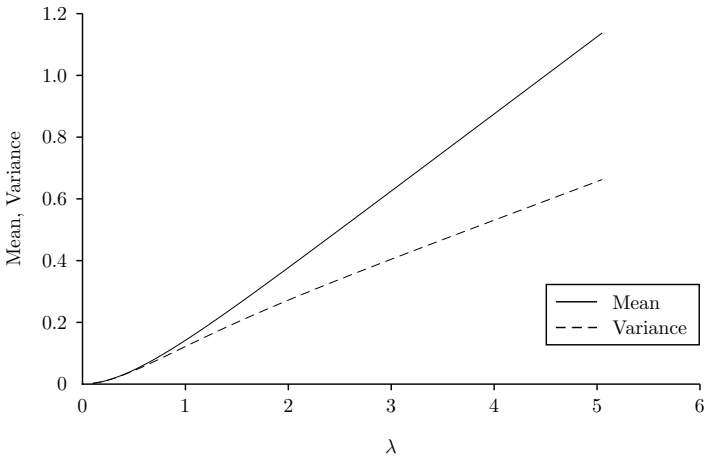
Alternatively, we can model the blockage duration itself as a random variable. An example from labor economics would be the number of unemployment episodes. A second unemployment episode requires termination of the first one, and hence, the time of blockage can be thought of as random.

A manageable result can be obtained under the following assumptions: Let the events arrive according to a Poisson process, and assume that the blockage time is independently distributed with exponential density function. Since the Poisson process is memoryless, the waiting time from the moment where blockage ends until the next occurrence is also exponentially distributed. Therefore, the distribution of the interarrival time between two events is a convolution of two exponential distributions. If we assume, admittedly unrealistically, that both distributions share a common parameter λ, the interarrival time has a gamma distribution with parameter $\alpha = 2$ (see Chap. 2.7.3). Moreover, the distribution of the number of events is of Erlangian type, with

$$f(y) = e^{-\lambda} \left(\frac{\lambda^{2y}}{(2y)!} + \frac{\lambda^{2y+1}}{(2y+1)!} \right) \quad y = 0, 1, \ldots \tag{3.21}$$

(See also 2.114). Fig. 3.2 plots the mean and variance of this distribution for $0.1 < \lambda < 5$. As for fixed blockage, this distribution is under-dispersed. Clearly, it would be desirable in future work to lift the restriction that the parameters of the two exponential distributions be the same.

Fig. 3.2. Mean and Variance of Exponential Blockage Model for $0.1 < \lambda < 5$



3.2 Maximum Likelihood Estimation

3.2.1 Introduction

This section is concerned with the problem of estimating β, the $(k \times 1)$ vector of regression coefficients in the Poisson regression model. Most of this chapter deals with the maximum likelihood method as it is the most common method to estimate count data models. The maximum likelihood principle states that the parameter should be chosen as to maximize the probability that the specified model has generated the observed sample. Numerous good econometric references to the general principles maximum likelihood estimation are available, including Amemiya (1985) and Cramer (1986).

3.2.2 Likelihood Function and Maximization

Given an independent sample of n pairs of observations (y_i, x_i), the joint probability distribution of the sample is the product of the individual conditional probability distributions:

$$f(y_1, \ldots, y_n | x_1, \ldots, x_n; \beta) = \prod_{i=1}^{n} f(y_i | x_i; \beta) \tag{3.22}$$

Understood as a function of the parameters, (3.22) is called *likelihood function*, and we write

$$L = L(\beta; y_1, \ldots, y_n, x_1, \ldots, x_n) \tag{3.23}$$

The maximum likelihood estimator is defined as

$$\hat{\beta} = \underset{\beta}{\operatorname{argmax}}\ L(\beta; y_1, \ldots, y_n, x_1, \ldots, x_n)$$

As the logarithm is a monotonic transformation, maximization of the likelihood function is equivalent to maximization of the logarithmic, or log-likelihood function $\ell = \log L$. In general, this transformation simplifies matters considerably, as it replaces products by sums. Moreover, it allows the use of the central limit theorem when studying the properties of the maximum likelihood estimator. The log-likelihood function for the Poisson regression model takes the form

$$\begin{aligned} \ell(\beta; Y, X) &= \log \prod_{i=1}^{n} f(y_i | x_i; \beta) \\ &= \sum_{i=1}^{n} \log f(y_i | x_i; \beta) \\ &= \sum_{i=1}^{n} -\exp(x_i'\beta) + y_i x_i'\beta - \log(y_i!) \end{aligned} \tag{3.24}$$

The maximizing value for β, denoted as $\hat{\beta}$, is found by computing the first derivatives of the log-likelihood function and setting them equal to zero. In the Poisson regression model, there are k such derivatives, with respect to β_1, β_2 and so forth. The (column) vector that collects these k first derivatives is alternatively denoted as *gradient vector*, or as *score vector*. The latter term is used in the following. We write

$$s_n(\beta; y, x) = \frac{\partial \ell(\beta; y, x)}{\partial \beta} = \sum_{i=1}^{n} [y_i - \exp(x_i'\beta)] x_i$$

We use the subscript "n" as a reminder that the score depends on the sample size. The maximum likelihood estimator $\hat{\beta}$ is the value of β that solves the first order conditions for a maximum

$$s_n(\hat{\beta}; y, x) = 0 \tag{3.25}$$

Note that as long as x_i includes a constant, the first order conditions imply that $\sum_{i=1}^{n} \hat{u}_i = 0$, where $\hat{u}_i$ is an implicit residual defined as

$$\hat{u}_i = y_i - \hat{\mathrm{E}}(y_i|x_i) = y_i - \exp(x_i'\hat{\beta})$$

For non-constant regressors, (3.25) can be interpreted as a set of orthogonality conditions:

$$\sum_{i=1}^{n} \hat{u}_i x_j = 0,\ j = 2, \ldots, k$$

Equation (3.25) gives the necessary conditions for a maximum. If, in addition, the matrix of second derivatives, the Hessian matrix, is negative definite for all values of β, the solution to (3.25) is called the *maximum likelihood estimator*. The Hessian matrix of the Poisson log-likelihood function is given by

$$\begin{aligned} H_n(\beta; y, x) &= \frac{\partial^2 \ell(\beta; y, x)}{\partial \beta \partial \beta'} \\ &= -\sum_{i=1}^{n} \exp(x_i'\beta) x_i x_i' \end{aligned} \tag{3.26}$$

H_n is negative definite, the log-likelihood function of the Poisson regression model is globally concave, and the set of parameters solving the first-order conditions are the unique maximum likelihood estimators.

3.2.3 Newton-Raphson Algorithm

Since (3.25) is non-linear in β, the system of k equations has to be solved using an iterative algorithm. A common choice that works well for concave objective functions is the Newton-Raphson method. It can be motivated as follows. Given any initial parameter estimate, say $\hat{\beta}^0$, we can obtain a second-order approximation of $\ell(\beta)$ around $\hat{\beta}^0$:

$$\ell^*(\beta) = \ell(\hat{\beta}^0) + s_n(\hat{\beta}^0)'(\beta - \hat{\beta}^0) + \frac{1}{2}(\beta - \hat{\beta}^0)' H_n(\hat{\beta}^0)(\beta - \hat{\beta}^0) \approx \ell(\beta)$$

Now, we can maximize $\ell^*(\beta)$ (rather than $\ell(\beta)$) with respect to β, yielding a new parameter value which we call $\hat{\beta}^1$. The first order condition of this simpler problem is

$$s_n(\hat{\beta}^0) + H_n(\hat{\beta}^0)(\hat{\beta}^1 - \hat{\beta}^0) = 0$$

or

$$\hat{\beta}^1 = \hat{\beta}^0 - [H_n(\hat{\beta}^0)]^{-1} s_n(\hat{\beta}^0)$$

Thus, for arbitrary starting value $\hat{\beta}^0$, the Newton-Raphson updating rule is given by

$$\hat{\beta}^{t+1} = \hat{\beta}^t - [H_n(\hat{\beta}^t)]^{-1} s_n(\hat{\beta}^t) \qquad t = 0, 1, \ldots \tag{3.27}$$

where $s(\cdot)$ denotes the score and $H(\cdot)$ the Hessian of the Poisson log-likelihood function. If we evaluate the right hand side at the maximum likelihood estimator, we observe that $s(\hat{\beta}^t) = 0$ and therefore $\hat{\beta}^{t+1} = \hat{\beta}^t$.

The iterative procedure ends when a predefined convergence criterion is satisfied. Possible criteria include the change in the value of the estimate $\hat{\beta}^{t+1} - \hat{\beta}^t$, the change in the log-likelihood $\ell(\hat{\beta}^{t+1}) - \ell(\hat{\beta}^t)$, or the value of the score at the estimate $s(\hat{\beta}^t)$. Convergence occurs when any of these values, or a combination of them, are close to zero (say, smaller than 10^{-5} in absolute value).

Numerical Derivatives

The algorithm presented in (3.27) could be based on analytical first and second derivatives of the log-likelihood function. A common alternative is the use of numerical derivatives. Numerical derivatives are the preferred option whenever analytical derivatives are difficult to establish. But even in cases such as the Poisson regression model, where the derivation of score and Hessian is relatively simple, numerical derivatives may lower the risk of programming errors. A downside is that numerical derivatives are considerably more time consuming. However, progress in computing speed has reduced the importance of this limitation, unless data sets and the number of parameters are very large (for instance, calculation time for a numerical Hessian is a quadratic function of the size of the matrix).

The standard formulas for numerical derivatives are $[f(b_i + h_i) - f(b_i)]/h_i$ if forward calculation is chosen, or $[f(b_i + h_i/2) - f(b_i - h_i/2)]/h_i$ if centered calculation is used. Methods differ in the way they determine h_i. For instance, StataCorp. (1997) uses an algorithm where h_i is selected such that $\varepsilon_1(|f(b_i)| + \varepsilon_1) \leq |f(b_i + h) - f(b_i)| \leq \varepsilon_2(|f(b_i)| + \varepsilon_2)$ for $\varepsilon_1 < \varepsilon_2$.

3.2.4 Properties of the Maximum Likelihood Estimator

The maximum likelihood estimator $\hat{\beta} = \text{argmax}\, L(\beta)$ is in general a non-linear function of the dependent variable. Therefore, analytical results on the small sample properties of the sampling distribution of $\hat{\beta}$ are unavailable. Provided a number of regularity conditions are satisfied, it can be shown that the maximum likelihood estimator is:

- asymptotically unbiased
- asymptotically normal
- asymptotically efficient

These three observations are summarized in the following convergence result (see Cramer, 1986, Amemiya, 1985):

$$\sqrt{n}(\hat{\beta}_{\text{ML}} - \beta_0) \xrightarrow{d} N\left(0, I(\beta_0)^{-1}\right) \tag{3.28}$$

where $\xrightarrow{d}$ stands for "converges in distribution", and where the Fisher information matrix $I(\beta_0)$ equals minus the expected value of the Hessian matrix of an observation evaluated at the true parameter vector β_0:

$$I(\beta_0) = -\text{E}\left[\frac{\partial^2 \ell(\beta; y_i, x_i)}{\partial \beta \partial \beta'}\right]_{\beta_0} \tag{3.29}$$

The maximum likelihood estimator is asymptotically unbiased (and, because of mean squared error convergence, consistent) since the distribution it converges to is centered at the true parameter value β_0. It is asymptotically efficient, since its variance is equal to the inverse of the Fisher information, the Cramér-Rao lower bound for any unbiased estimator.

While these asymptotic properties in a strict sense only hold in the limit of infinite sample size, in practice they are often assumed to be approximately valid, especially when the sample size is not that small. The approximate distribution of $\hat{\beta}$ is then given by

$$\hat{\beta}_{\text{ML}} \overset{\text{app}}{\sim} N(\beta_0, [nI(\beta_0)]^{-1}) \tag{3.30}$$

This result requires, in general, that the model is correctly specified. Let the true (conditional) density be denoted as $f_0(y_i|x_i)$. There must exist a β_0 such that

$$\prod_{i=1}^{n} f(y_i|x_i; \beta_0) = \prod_{i=1}^{n} f_0(y_i|x_i) \tag{3.31}$$

Properties of maximum likelihood estimation in misspecified models are discussed in the next section. Apart from correct specification, some further regularity conditions are required, essentially in order to ensure that the operations of differentiation and taking expectations can be interchanged. The first and second derivatives of the log-likelihood function must be defined, and the Fisher information matrix must be non-zero (see, for instance, Cramer, 1986, for further details).

Sketch of a Proof

The main steps of a proof are as follows. As starting point, consider a first-order Taylor series approximation of $s_n(\hat{\beta})$ around the true parameter vector β_0:

$$s_n(\hat{\beta}_n) \approx s_n(\beta_0) + H_n(\beta_0)(\hat{\beta}_n - \beta_0)$$

Since $s_n(\hat{\beta}_n)$ is zero by definition of a maximum likelihood estimator, we have

$$\hat{\beta}_n - \beta_0 \approx -H_n(\beta_0)^{-1} s_n(\beta_0)$$

or

$$\sqrt{n}(\hat{\beta}_n - \beta_0) \approx \left(-\frac{1}{n} H_n(\beta_0)\right)^{-1} \frac{1}{\sqrt{n}} s_n(\beta_0)$$

Now on one hand,

$$-\frac{1}{n} H_n(\beta_0) = -\frac{1}{n} \sum_{i=1}^{n} H_i(\beta_0)$$

converges almost surely to its first moment by the law of large numbers:

$$\lim_{n\to\infty} -\frac{1}{n} \sum_{i=1}^{n} H_i(\beta_0) = -\mathrm{E}(H(\beta_0)) = I(\beta_0)$$

On the other hand

$$\frac{1}{\sqrt{n}} s_n(\beta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[\frac{\partial \log f(y_i|x_i;\beta)}{\partial \beta}\right]_{\beta_0} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} s_i(\beta_0)$$

converges in distribution to a Normal distribution

$$\mathrm{N}(0, I(\beta_0))$$

by the Central Limit Theorem. The variance of the limit distribution follows from the information matrix equality,

$$\mathrm{Var}(s(\beta_0)) = -\mathrm{E}(H(\beta_0) = I(\beta_0)$$

Premultiplying by $I(\beta_0)^{-1}$, we obtain

$$\sqrt{n}(\hat{\beta}_n - \beta_0) \approx \left(-\frac{1}{n} H_n(\beta_0)\right)^{-1} \frac{1}{\sqrt{n}} s_n(\beta_0) \xrightarrow{d} N(0, I(\beta_0)^{-1})$$

which was to be shown (see 3.28).

3.2.5 Estimation of the Variance Matrix

For the purpose of inference, we will need to compute $\text{Var}(\hat{\beta}) = [nI(\beta_0)]^{-1}$, where $I(\beta_0)$ depends on the unknown β_0. While $\text{Var}(\hat{\beta})$ is thus unknown in principle, we can base inference on an estimated variance matrix. There are a number of consistent, and thus asymptotically equivalent, estimators. All of them rely on replacing the unknown β_0 by the consistent maximum likelihood estimator estimator $\hat{\beta}$.

The expected Hessian matrix can be computed in principle, which would lead to the first possible variance estimator

$$\widehat{\text{Var}}(\hat{\beta})_1 = -[\text{E}H_n(\hat{\beta})]^{-1}$$

It is almost always the case, however, that the Hessian matrix is a highly nonlinear function of x and y, making it practically impossible to take expectations in all but the most trivial cases (the constant only model below is one such example). Similarily, straight calculation of the variance of the score is likely to fail.

More importantly, in conditional probability models, we do not even specify the marginal distribution of x, so that the unconditional expectation cannot be taken. In practice, therefore, we need to estimate the information matrix by using sample means rather than expectations, which is straigtforward to do. With $\widehat{I(\beta_0)} = -n^{-1}H_n(\hat{\beta})$, we obtain

$$\widehat{\text{Var}}(\hat{\beta})_2 = -[H_n(\hat{\beta})]^{-1}$$

Based on the sample variance of the score (the summed outer product of the score), and thus exploiting the information matrix equality, we obtain

$$\widehat{\text{Var}}(\hat{\beta})_3 = \left[\sum_{i=1}^{n} s_i(\hat{\beta})s_i(\hat{\beta})'\right]^{-1}$$

Example: Constant-Only Poisson Model

If the only regressor is a constant, we find that

$$s_n = \sum_{i=1}^{n} -1 + \frac{y_i}{\lambda}$$

$$H_n = \sum_{i=1}^{n} -\frac{y_i}{\lambda^2}$$

From the first order condition the ML estimator is given by $\hat{\lambda} = \bar{y}$. In this case, the ML estimator coincides with least squares and method of moments estimators. The approximate distribution of the ML estimator in the constant-only Poisson model is

$$\hat{\lambda}_{\mathrm{ML}} \overset{\text{app}}{\sim} N\left(\beta_0, \widehat{\mathrm{Var}}(\hat{\lambda}_{\mathrm{ML}})\right) \tag{3.32}$$

where $\widehat{\mathrm{Var}}(\hat{\beta}_{\mathrm{ML}}) = [n\widehat{I(\beta_0)}]^{-1}$ can computed by either of the three methods.

Expected Hessian.

$$\widehat{\mathrm{Var}}_1(\hat{\lambda}) = \left[-n\mathrm{E}\left[\frac{y_i}{\lambda^2}\right]_{\hat{\lambda}}\right]^{-1} = \frac{\hat{\lambda}}{n}$$

Actual Hessian.

$$\widehat{\mathrm{Var}}_2(\hat{\lambda}) = -\left[\sum_{i=1}^{n} -\frac{y_i}{\lambda^2}\right]^{-1}_{\hat{\lambda}} = \left[\frac{\sum_{i=1}^{n} y_i}{\lambda^2}\right]^{-1}_{\hat{\lambda}} = \frac{\hat{\lambda}}{n}$$

Outer Product of Score.

$$\widehat{\mathrm{Var}}_3(\hat{\lambda}) = \left[\sum_{i=1}^{n}\left(\frac{y_i}{\lambda} - 1\right)^2\right]^{-1}_{\hat{\lambda}} = \left[\sum_{i=1}^{n}\frac{(y_i-\lambda)^2}{\lambda^2}\right]^{-1}_{\hat{\lambda}} = \frac{\hat{\lambda}^2}{n\widehat{\mathrm{Var}}(y)}$$

Thus, the first two methods yield exactly the same variance estimators. The estimator based on the third method is different. However, it is asymptotically equivalent as long the model is correctly specified, since $\mathrm{Var}(y) = \lambda$ in this case. In general, these results are not surprising, since we know that for *i.i.d.* sampling from any distribution $\bar{y} \overset{\text{app}}{\sim} N(\mathrm{E}(y), \mathrm{Var}(y)/n)$.

For the constant-only model, it is also simple to compute the exact variance of the score (rather than its estimate, as in the outer product of the score formula), using again the fact that it is the case under the Poisson assumption that $\mathrm{Var}(y_i) = \mathrm{E}(y_i) = \lambda$, and therefore

$$\widehat{\mathrm{Var}}_4(\hat{\lambda}) = \left[\mathrm{Var}\sum_{i=1}^{n}\left(\frac{y_i}{\lambda} - 1\right)\right]^{-1} = \left[\frac{n\mathrm{Var}(y_i)}{\lambda^2}\right]^{-1}_{\hat{\lambda}} = \frac{\hat{\lambda}}{n}$$

a manifestation of the information matrix equality.

As we will see in the next Chapter 3.3 on pseudo-maximum likelihood, a violation of the information matrix equality is a key feature of misspecified maximum likelihood estimators, and in this case, none of the four covariance matrix estimators discussed here is valid.

3.2.6 Approximate Distribution of the Poisson Regression Coefficients

Returning to the case of a correctly specified Poisson model, the above results easily extend to the regression case with explanatory variables. As before

$$\hat{\beta} \overset{\text{app}}{\sim} N\left(\beta_0, [nI(\beta_0)]^{-1}\right) \tag{3.33}$$

but in this case

$$I(\beta_0) = -\mathrm{E}(H(\beta_0)) = \mathrm{E}\left[\exp(x'\beta_0)xx'\right]$$

Iterating expectations, we can write

$$\mathrm{E}\left[\exp(x'\beta_0)xx'\right] = \mathrm{E}_x\mathrm{E}_{y|x}\left[\exp(x'\beta_0)xx'\right] = \mathrm{E}_x\left[\exp(x'\beta_0)xx'\right]$$

since the Hessian does not depend on y. A natural estimator for the Fisher information is therefore

$$\widehat{I(\beta_0)} = \frac{1}{n}\sum_{i=1}^{n}\exp(x_i'\hat{\beta})x_ix_i'$$

where the unknown true parameter β_0 has been replaced by the consistent estimator $\hat{\beta}$, leading to the variance estimator

$$\widehat{\mathrm{Var}}_1(\hat{\beta}) = \left[\sum_{i=1}^{n}\exp(x_i'\hat{\beta})x_ix_i'\right]^{-1} \tag{3.34}$$

Similarly, one could estimate the variance of $\hat{\beta}$ by summing over the outer product of the score

$$\widehat{\mathrm{Var}}_2(\hat{\beta}) = \left[\sum_{i=1}^{n}(y_i - \exp(x_i'\hat{\beta}))^2x_ix_i'\right]^{-1} \tag{3.35}$$

In finite samples, $\widehat{\mathrm{Var}}_1$ and $\widehat{\mathrm{Var}}_2$ generally differ. Asymptotically, the two are the same provided the model is correctly specified because then $\mathrm{Var}(y_i|x_i) = \mathrm{E}(y_i - \exp(x_i'\beta_0)|x)^2$ and $\mathrm{E}(y_i|x_i) = \exp(x_i'\beta_0)$ are the same.

3.2.7 Bias Reduction Techniques

> "The few finite sample corrections that have been proposed remain extensively ignored by empiricists (...)" (Maasoumi, 1992, p. 2)

While the above statement by Maasoumi (1992) may be still valid, this should not prevent us from at least pointing to the problem and discussing some of the solution concepts that have been proposed in the literature. The problem is that the maximum likelihood technique, while providing a *consistent* estimator of the model parameters (if based on a correctly specified model), does not provide an *unbiased* estimator. For large n the probability of the estimates being 'close' to the true value gets larger and larger, but in small samples, there can be a bias of arbitrary magnitude. . In general, small sample bias arises because the score function $s(\beta)$ is (in most cases) non-linear in β.

This section is concerned with one part of that bias, the so-called first order bias that is at most of order $O(n^{-1})$, i.e. $n \times$ bias is bounded in probability for $n \to \infty$. We discuss methods that remove such bias. To understand the line of the argument, recall some basic theoretical properties of the score vector:

1. The score $s(\beta)$ is a random variable, since it depends on y. It was shown before that

   $$\mathrm{E}(s(\beta))|_{\beta=\beta_0} = 0$$

   where β_0 is the true parameter.

2. In an identified model, the log likelihood function is concave at the true parameter value, i.e., $\partial s(\beta)/\partial\beta'$ is negative definite. Let $s_j(\beta)$ denote the j-th element of the score vector. Then it must be the case that $\partial s_j(\beta)/\partial\beta_j < 0$.

3. The score function can be linear ($\partial^2 s_j(\beta)/(\partial\beta_k)^2 = 0$ for all j, k), convex ($\partial^2 s_j(\beta)/(\partial\beta_k)^2 > 0$ for some j, k), or concave ($\partial^2 s_j(\beta)/(\partial\beta_k)^2 < 0$ for some j, k).

If $s(\beta)$ is non-linear in β then, for finite n,

- $\mathrm{E}(\hat{\beta}) \geq \beta_0$ if $s(\beta)$ is convex, and
- $\mathrm{E}(\hat{\beta}) \leq \beta_0$ if $s(\beta)$ is concave,

where $\hat{\beta}$ denotes the value that solves the equation $s(\beta) = 0$, i.e., $\hat{\beta}$ is the maximum likelihood estimator, and where equality only holds in the limit, as the sample size approaches infinity.

As an illustration for these small sample biases, consider the following example, a Poisson model with single parameter $\lambda = \exp(\beta)$. The goal is to estimate β. For a sample of n independent observations, the score function is given by

$$s(\beta) = -n\exp(\beta) + \sum_{i=1}^{n} y_i \tag{3.36}$$

Moreover, the score function is concave, since The expected score at $\beta = \beta_0$ is zero, as it should, since

$$\mathrm{E}[s(\beta_0)] = -n\exp(\beta_0) + \mathrm{E}\left[\sum_{i=1}^{n} y_i\right] = -n\exp(\beta_0) + n\exp(\beta_0) = 0$$

$\partial^2 s(\beta)/(\partial\beta)^2 = -n\exp(\beta) < 0$. Solving (3.36) for β, we obtain the maximum likelihood estimator

$$\hat{\beta} = \log\bar{y}$$

where $\bar{y}$ is the sample mean of the data. From Jensen's inequality, we know that

$$\mathrm{E}[\hat{\beta}] = \mathrm{E}[\log\bar{y}] \leq \log\mathrm{E}[\bar{y}] = \beta_0$$

which was to be shown. In finite samples, the maximum likelihood estimator for β_0 in this model is downward biased. The bias disappears asymptotically

– the maximum likelihood estimator is consistent – because it depends on a positive variance of $\bar{y}$, which goes to zero at rate $1/n$. Note that the small sample properties depend on the parameterization. By parameterizing the mean directly as $\lambda = \beta$, a linear score function is obtained, and the estimator is unbiased.

Unfortunately, such a re-parametrization is only possible in the simple Poisson model without regressors. The score of the Poisson regression model with regressors is in general inherently non-linear

$$s(\beta) = \sum_{i=1}^{n} (y_i - \exp(x_i'\beta))x_i$$

and the Poisson maximum likelihood estimator is therefore biased in finite samples. Whether the score function is convex or concave depends on the values of x. This section discusses a method that removes the first order bias $b(\beta)$ from the Poisson estimates. The first order bias may implicitly be defined as

$$\mathrm{E}(\hat{\beta} - \beta) = b(\beta) + O(n^{-2})$$

where $O(n^{-2})$ denotes terms that are at most of order in probability n^{-2}, i.e. converge to zero at rate n^{δ} where $\delta > -2$. The first order bias $b(\beta)$ has to be calculated depending on the model under investigation. Assume that this has been done. Then, there are two approaches for removing the first order bias from the maximum likelihood estimator. First, $\hat{\beta}$ may be calculated as usual, and the bias is subtracted after estimation (see, for instance, McCullagh and Nelder, 1989, Chap. 15.2). Alternatively, the bias may be removed by artificially introducing a bias in the score equation (Firth, 1992). Consider the following modified score equation:

$$s^*(\beta) = s(\beta) - I(\beta)b(\beta) \tag{3.37}$$

where I is the Fisher information. Taking roots of $s^*(\beta)$, instead of $s(\beta)$, yields an estimator β^*. In general, unless $b(\beta) = 0$, this estimator is not equal to the maximum likelihood estimator $\hat{\beta}$. First note that from (3.37)

$$s^*(\hat{\beta}) = -I(\hat{\beta})b(\hat{\beta}) \tag{3.38}$$

Expanding s^* around $\hat{\beta}$ and evaluating at $\beta = \beta^*$ yields:

$$s^*(\beta^*) \approx s^*(\hat{\beta}) + H^*(\hat{\beta})(\beta^* - \hat{\beta})$$

But the left hand side is zero by definition. Further, $H^*(\beta) = -I(\beta) + O(n^{-1})$. Finally, using (3.38)

$$-I(\hat{\beta})b(\hat{\beta}) - I(\hat{\beta})(\beta^* - \hat{\beta}) \approx 0$$

or

$$\beta^* \approx \hat{\beta} - b(\hat{\beta})$$

Thus, introducing a weighted first order bias term in the score function gives bias corrected roots, i.e., bias corrected estimates.

Firth (1992) shows that, for the log-linear Poisson model, the bias-corrected score can be written as

$$s^*(\beta; y, x) = \sum_{i=1}^{n} (y_i + h_i/2 - \exp(x_i'\beta))x_i' \tag{3.39}$$

where h_i is the i-th diagonal element of the matrix $H = WX(X'WX)^{-1}X'$, W is an $(n \times n)$ matrix having the individual variances on its main diagonal and zeros elsewhere, and X is the $(n \times k)$ matrix of regressors.

The following Monte Carlo experiments illustrate the existence, and the consequences of the removal of the first order bias. Four different sample sizes were considered ($n = 10, 20, 50$, and 100). For each sample size, 1000 vectors of y were drawn from a Poisson distribution with fixed mean-vector λ. λ was constructed as $\lambda_i = -1 + x_i$, where x_i are independent drawings from a normal distribution with mean 0 and variance 4. The true slope coefficient, on which the discussion will focus, is thus $\beta_0 = 1$. For each of the 4×1000 datasets, a Poisson regression with and without bias correction was performed. Some characteristics of the empirical distributions of $\hat{\beta}$ and β^* are given in Tab. 3.1 Except for n=100, the mean of β^* is closer to the true value than the mean of the MLE $\hat{\beta}$. Always, the maximal deviations from the true value in both directions are smaller for β^*. The bias corrected estimator has also a smaller standard error. Thus, the Monte Carlo evidence suggests that using the bias reduction provides some gain in bias reduction as well as in precision, in particular when the sample size is moderate. The dependence of the effect of the correction on sample size does not come as a surprise, since we know that the Poisson maximum likelihood estimator is consistent in this set-up.

3.3 Pseudo-Maximum Likelihood

White (1982) considers a situation where the model is not correctly specified: There exists no β such that

$$\sum_{i=1}^{n} \log[f(y_i|x_i, \beta)] = \sum_{i=1}^{n} \log[f_0(y_i|x_i)] \tag{3.40}$$

where f is the specified conditional density and $f_0(y|x)$ is the true conditional density. Obtaining parameter estimates by maximizing a misspecified log-likelihood

$$\ell(\beta; y, x) = \sum_{i=1}^{n} \log[f(y_i|x_i, \beta)]$$

Table 3.1. Bias Reduced Poisson Estimates

		ML-Poisson Estimates $\hat{\beta}$	Bias Reduced Estimates β^*
n=10	Mean	1.0290	1.0029
	Standard Error	0.2333	0.2222
	Minimum	0.4027	0.3815
	Maximum	2.4715	2.1479
n=20	Mean	1.0357	1.0062
	Standard Error	0.1712	0.1578
	Minimum	0.6158	0.6069
	Maximum	1.7512	1.6238
n=50	Mean	1.0022	0.9996
	Standard Error	0.0601	0.0598
	Minimum	0.8366	0.8349
	Maximum	1.2251	1.2212
n=100	Mean	0.9984	0.9960
	Standard Error	0.0487	0.0483
	Minimum	0.8062	0.8050
	Maximum	1.1515	1.1475

is a method that is conventionally referred to as *quasi maximum likelihood estimation* (QML).

The consequences of misspecification can be best stated considering the asymptotic distribution of a quasi maximum likelihood estimator (QMLE). Under the assumption of independent sampling, the QMLE has in fact a well defined limiting distribution given by (See White, 1982)

$$\sqrt{n}(\tilde{\beta} - \beta^*) \xrightarrow{d} N(0, I^{-1} J I^{-1}) \tag{3.41}$$

The quasi maximum likelihood estimator $\tilde{\beta}$ is a consistent estimator for a pseudo-true value β^*, where β^* minimizes the Kullback-distance between the specified model and the true model:

$$\mathcal{K}\{f_0(y|x), f(y|x;\beta)\} = \mathrm{E}_0\left[\log \frac{f_0(y|x)}{f(y|x;\beta)}\right] \tag{3.42}$$

The asymptotic covariance is given by $I^{-1} J I^{-1}$, where I is minus the expected Hessian of an observation, as before,

$$I = -\mathrm{E}(H(\beta_0))$$

and J is the variance of the score,

$$J = \mathrm{Var}(s(\beta_0))$$

If the model is correctly specified, $\beta^* = \beta_0$ and $I = J$. Thus, the limiting distribution of the QML collapses to the limiting distribution of the maximum likelihood estimator with covariance matrix $I^{-1} = J^{-1}$.

Under misspecification, little can be said about the relationship between the QMLE $\tilde{\beta}$ and the maximum likelihood estimator $\hat{\beta}$. The two main criteria for 'good' estimators are consistency and efficiency. The QMLE is in general inconsistent (for β_0) and inefficient. However, Gourieroux, Monfort and Trognon (1984a) give a condition under which $\tilde{\beta} \xrightarrow{p} \beta_0$. It refers to the special case, where the mean is correctly specified, i.e., there exists a $\beta = \beta_0$ such that

$$\mu(\beta_0, x) = \mu_0(x),$$

it holds that β is *first order identifiable*, i.e., $\mu(x, \beta) = \mu(x, \beta_0)\ \forall x$ implies that $\beta = \beta_0$, and the quasi-likelihood function f is a member of the linear exponential family. Gourieroux, Monfort and Trognon (1984a) refer to this situation as *pseudo maximum likelihood estimation* (PMLE). They show that a quasi maximum likelihood estimator is consistent for β_0 if and only if the distribution family of the estimated model is a linear exponential family (and thus QML=PML).

3.3.1 Linear Exponential Families

Linear exponential families can be written in the form

$$f(y, m) = \exp\{A(m) + B(y) + C(m)y\}$$

where m is the mean of the distribution. For example, the Poisson distribution is a linear exponential family with $A(m) = -m$, $B(y) = -\log y!$, and $C(m) = \log(m)$. Similarly, the normal distribution is a linear exponential family, since we can write

$$\exp\left\{-\frac{1}{2}\left(\frac{y-m}{\sigma}\right)^2\right\} = \exp\left\{-\frac{1}{2}\frac{y^2}{\sigma^2} + \frac{y\,m}{\sigma} - \frac{m^2}{\sigma^2}\right\}$$

Linear exponential families have the property that

$$\frac{\partial A(m)}{\partial m} + \frac{\partial C(m)}{\partial m} m = 0 \tag{3.43}$$

which implies that

$$m = -\frac{\partial A(m)/\partial m}{\partial C(m)/\partial m} \tag{3.44}$$

This property follows from differentiation of the identity (in the case of a discrete random variable, we would need to replace the integral by an appropriately defined sum)

$$\int f(y,m)dy = 1$$

with respect to m, yielding

$$\frac{d}{dm}\int f(y,m)dy = 0$$

We can change the order of differentiation and integration and therefore rewrite the left side of the equation as

$$\begin{aligned}\int \frac{d\log f(y,m)}{dm} f(y,m)dy &= \int \left(\frac{\partial A(m)}{\partial m} + \frac{\partial C(m)}{\partial m}y\right) f(y,m)dy \\ &= \mathrm{E}\left(\frac{\partial A(m)}{\partial m} + \frac{\partial C(m)}{\partial m}y\right) \\ &= \frac{\partial A(m)}{\partial m} + \frac{\partial C(m)}{\partial m}m\end{aligned}$$

which establishes property (3.43).

Now, the log likelihood of an observation from a linear exponential family is

$$\log f(y,m) = A(m) + B(y) + C(m)y$$

with score function

$$\frac{\partial \log f(y,m)}{\partial m} = \frac{\partial A(m)}{\partial m} + \frac{\partial C(m)}{\partial m}y$$

which we can rewrite as

$$\frac{\partial C(m)}{\partial m}\left(\frac{\partial A(m)/\partial m}{\partial C(m)/\partial m} + y\right) = \frac{\partial C(m)}{\partial m}(y-m)$$

where we have used relationship (3.44). Thus, the standard result that the expected score is equal to zero is, for linear exponential families, equivalent to saying that $\mathrm{E}(y) = m$. Since, when estimating m using maximum likelihood, the actual score converges against the expected score, the MLE is consistent provided the mean is correctly specified. Correct specification becomes an issue as soon as m is parameterized in terms of regressors, for instance when $m(x) = \exp(x'\beta)$.

3.3.2 Biased Poisson Maximum Likelihood Inference

Since the Poisson distribution is a linear exponential family, we can conclude that the maximum likelihood estimator $\hat{\beta}$ is consistent as long as the conditional expectation is correctly specified, regardless of whether or not the true data generating process is really a Poisson distribution. If it is not – if there are departures from the Poisson specification in higher order moments – the estimator is the Pseudo Maximum Likelihood estimator.

The leading cause of departure from the Poisson assumption is overdispersion, a conditional variance that exceeds the conditional mean. If such overdispersion is present then, according to the Gourieroux, Monfort and Trognon (1984a) result, one can continue using the Poisson model, but one should base inference on the asymptotic PML variance matrix

$$n^{-1}I^{-1}JI^{-1}$$
$$= n^{-1}\mathrm{E}\left[\exp(x'\beta_0)xx'\right]^{-1}\mathrm{E}\left[(y-\exp(x'\beta_0))^2xx'\right]\mathrm{E}\left[\exp(x'\beta_0)xx'\right]^{-1}$$

rather than the asymptotic ML variance matrix

$$n^{-1}I^{-1} = n^{-1}\mathrm{E}\left[\exp(x'\beta_0)xx'\right]^{-1}$$

In other words, using the Poisson regression model for a non-Poisson population estimates the right parameter values, on average (consistency) but gets the inference wrong. Conventional Wald tests do not have the right size.

Interestingly, the direction of the bias can be established in some cases, as there is a direct connection between the departure from equidispersion, i.e., overdispersion or underdispersion, and the direction of bias when erroneously using the Poisson ML variance matrix. In order to compute the direction of the bias, we need to establish the matrix difference

$$I^{-1} - I^{-1}JI^{-1} = I^{-1}(I-J)I^{-1}$$

But $I - J$ is readily obtained as

$$I - J = \mathrm{E}\left[\exp(x'\beta_0)xx'\right] - \mathrm{E}\left[(y-\exp(x'\beta_0))^2xx'\right]$$
$$= \mathrm{E}_x[\mathrm{E}(y|x)xx'] - \mathrm{E}_x[\mathrm{Var}(y|x)xx'] = \mathrm{E}_x[(\mathrm{E}(y|x) - \mathrm{Var}(y|x))xx']$$

This means that if $\mathrm{E}(y|x) < \mathrm{Var}(y|x)$ (overdispersion), the difference is negative, in a matrix sense, and we know therefore, that the ML variance matrix is smaller than the PML variance matrix, which, in this case, is the correct one. In other words, ignoring overdispersion and applying the standard variance estimator under the maximum likelihood assumption leads to an underestimation of the true standard errors. Spurious inference may result, as t-values will tend to be inflated. The opposite situation arises with underdispersion, a conditional variance smaller than the conditional mean. In this case, the ML variance matrix overestimates the true variance matrix, and t-values will tend to be too small.

3.3.3 Robust Poisson Regression

Of course, the problem can be easily avoided by applying PML standard errors that will lead to asymptotically valid inference. The merits of this "method", using the Poisson regression model together with the sandwich variance estimator are clear. As long as we are confident about the validity of our mean

function, we can remain largely agnostic with respect to higher order moments, apply Poisson regression, and obtain consistent parameter estimates as well as valid inference, at least for large enough samples.

In this section, the implications of the PML result are explored in the context of the Poisson regression model. PML estimation exploits the fact that, as the Poisson distribution is a linear exponential family, departure from the standard variance function does not affect consistency of the parameter estimates as long as the mean is correctly specified. The only effect of a misspecified variance function is then that the estimated variance matrix under the maximum likelihood assumption is "wrong" and has to be adjusted.

The approximate distribution of the Poisson PMLE in large but finite sample is

$$\hat{\beta} \overset{\text{app}}{\sim} N\left(\beta_0, \widehat{\text{Var}}(\hat{\beta})\right) \tag{3.45}$$

where

$$\widehat{\text{Var}}(\hat{\beta}) = \left[\sum_{i=1}^{n} x_i x_i' \exp(x_i'\hat{\beta})\right]^{-1} \sum_{i=1}^{n} x_i x_i' \widehat{\text{Var}}(y_i|x_i) \left[\sum_{i=1}^{n} x_i x_i' \exp(x_i'\beta)\right]^{-1}$$

and the population expressions $\text{E}(y_i|x_i)$ and $\text{Var}(y_i|x_i)$ have been replaced by their sample equivalents. As indicated above, *pseudo-maximum likelihood* estimation, or *robust Poisson regression* leads to consistent estimation of both parameters as well as standard errors.

The gist of this approach is very much like the use of Huber-White standard errors in the linear model. Here as there, the problem of heteroskedasticity, a violation of one of the standard assumptions of the linear model, calls for either one of two responses. We can either rely on consistent estimation of the regression coefficients and try to adjust the covariance matrix in order to obtain valid inference. This is the robust regression approach discussed here. Or we can attempt to model the heteroskedasticity directly in an attempt to use all information in the data efficiently, and thus obtain a more efficient estimator. In the linear model, this leads to the method of Generalized Least Squares. In both cases, there is a price to pay for attempting efficient estimation, namely the potential loss of consistency. Herein lies the general appeal of robust estimation strategies.

An implementation issue unanswered so far is the question how to estimate $\widehat{\text{Var}}(y_i|x_i;\hat{\beta})$ in (3.45). The most obvious approach, corresponding to the outer product of the score formula, would use

$$\widehat{\text{Var}}(y_i|x_i;\hat{\beta}) = (y_i - \exp(x_i'\hat{\beta}))^2$$

This is a direct application of the White (1982) result, see also Breslow (1990).

There have been some alternative proposals in the literature that put more structure on the variance function. For example, if we are willing to assume that the conditional variance is a linear function of the conditional mean, such that (McCullagh and Nelder, 1989)

$$\widehat{\mathrm{Var}}(y_i|x_i;\hat{\beta}) = \hat{\sigma}^2 \exp(x_i'\hat{\beta}),$$

the estimated variance matrix of $\hat{\beta}$ thus simplifies to

$$\widehat{\mathrm{Var}}(\hat{\beta}) = \hat{\sigma}^2 \left[\sum_{i=1}^{n} x_i x_i' \hat{\lambda}_i\right]^{-1}$$

where $\hat{\sigma}^2$ can be estimated using the moment estimator:

$$\hat{\sigma}^2 = \frac{1}{n-k} \sum_{i=1}^{n} \frac{(y_i - \hat{\lambda}_i)^2}{\hat{\lambda}_i}$$

If, on the other hand, one believes in a quadratic variance function (Gourieroux, Monfort and Trognon, 1984b)

$$\widehat{\mathrm{Var}}(y_i|x_i;\hat{\beta}) = \hat{\lambda}_i + \hat{\sigma}^2 \hat{\lambda}_i^2$$

an estimate for σ^2 can be obtained by the auxiliary regression (See also Cameron and Trivedi, 1990):

$$(y_i - \hat{\lambda}_i)^2 - \hat{\lambda}_i = \sigma^2 \hat{\lambda}_i^2 + v_i$$

This regression yields a strongly consistent estimator for σ^2.

A cautionary remark applies to the third assumption of a quadratic variance function. While for the first two assumptions, PML estimation based on the Poisson distribution uses the *available* information efficiently, this is not the case for the third. Gourieroux, Monfort and Trognon (1984b) show that incorporating this information on the variance into the estimation provides a gain in efficiency, and they call this procedure *quasi-generalized pseudo maximum likelihood estimation* (QGPML). In the special case, where the true density is itself a member of a linear exponential family (which is fully characterized by its first two moments), QGPML estimation is asymptotically equivalent to ML estimation and hence fully efficient.

Monte Carlo Study

Given the three different estimators for a robust variance covariance matrix of $\hat{\beta}$, a Monte Carlo study might give some indication on whether they provide a substantial gain as compared to the use of the conventional Poisson-variance matrix. We also study, whether the three estimators lead to substantially different results and whether the validity of inference is robust with respect to the choice of a particular assumption. The latter finding would increase the overall confidence in Robust Poisson regression since otherwise one would need to rely, for example, on pre-tests to justify the particular variance assumption.

There have been a number of previous Monte Carlo studies to evaluate the finite sample properties of the Poisson regression model with adjusted

covariance matrix estimator. Examples are Winkelmann and Zimmermann (1992a), Brännäs (1992a) and Bourlange and Doz (1988). The earlier results are extended by including an investigation of the performance of the variance estimator appropriate under the assumption of a quadratic variance function.

The design of the study is as follows: Firstly, random samples of count data with different degrees of overdispersion are drawn. Then, PMLE's are obtained based on the Poisson regression model. This procedure is repeated 1000 times. The sampling distribution of the PMLE over the 1000 replications can be investigated.

Simulating random counts with equidispersion proceeded as follows: A linear predictor is modeled as

$$\eta = -1 + x$$

where x is a $(n \times 1)$ vector of standard uniform random numbers. Thus, the true parameter vector is β_0 =(-1,1)'. The predictor is kept fixed over replications as we are interested in the *conditional* distribution of y_i given λ_i, where $\lambda_i = \exp(\eta_i)$ and $i = 1, \ldots, n$. Pseudo-random Poisson numbers are obtained with an algorithm described in Knuth (1969).

To obtain random counts with overdispersion, a stochastic expected value is introduced: $\tilde{\lambda}_i = \exp(\eta_i + \varepsilon_i) = \lambda_i u_i$. ε_i are independent random normal numbers with constant variance σ_ε^2 and mean $\mu_\varepsilon = -0.5\,\sigma_\varepsilon^2$. Thus $u_i = \exp(\varepsilon_i)$ has mean $\mathrm{E}(u) = 1$ and variance $\sigma_u^2 = \exp(\sigma_\varepsilon^2) - 1$. The conditional variance of the Poisson-log-normal model is given by $\mathrm{Var}(y_i|x_i) = \lambda_i + \sigma_u^2 \lambda_i^2$. The degree of overdispersion depends on σ_ε^2 which is chosen in a way as to yield the values $0.2, 1.0$ and 5.0 for σ_u^2. In this way, the experiments cover a range from modest to substantial overdispersion.

In order to study the impact of increasing sample size on the quality of the approximation for three degrees of overdispersion, the experiments were conducted for samples of size 100 and 1000, respectively. The results are given in Tab. 3.2 and Tab. 3.3. For both sample sizes the mean and standard errors of the slope coefficient $\hat{\beta}_1$ are given. Furthermore, the empirical size for a two-sided asymptotic t-test under two alternative nominal significance levels are reported. The t-values were calculated in four different ways, following the different possible assumptions discussed above: $\widehat{\mathrm{Var}}_{\mathrm{psn}}$ is based on the assumption of a correctly specified model. $\widehat{\mathrm{Var}}_{\mathrm{White}}$ allows for all kinds of misspecification, while $\widehat{\mathrm{Var}}_{\mathrm{lvf}}$ and $\widehat{\mathrm{Var}}_{\mathrm{qvf}}$ are based on specific violations of the variance assumption. In the former case, the variance is a linear function of the expected value, while in the latter case, it is assumed to be a quadratic function (this, incidentally, is the 'true' model).

In interpreting the results, the focus is on consistency and valid inference. For $n = 100$, the deviation of the mean of the estimated coefficient $\hat{\beta}_1$ from its true value 1.0 is only in the second decimal place. For $n = 1000$, the deviation reduces to the third decimal place. For both sample sizes, this holds independently of the magnitude of overdispersion. The sample standard error increases with increasing overdispersion, and decreases strongly with sample

size. These results confirm that the Poisson model yields consistent parameter estimates also in the presence of overdispersion.

A very different conclusion has to be drawn from the estimated standard errors. The empirical size of the test can be compared to its nominal size, i.e., the significance level of the test. If the empirical size exceeds the nominal size by an amount that is beyond the sampling variation expected from 1000 replications, it indicates an underestimation of the standard errors. And in fact, the test based on the Poisson variance estimator systematically overstates the nominal size. Not surprisingly, the underestimation of the true standard errors is the more severe, the larger the overdispersion. In the case of slight overdispersion, the asymptotic t-test for the smaller sample leads to type-I errors that are around 25% higher than the significance levels $\alpha = 0.05, 0.10$. For $\sigma_u^2 = 1$ and $n = 100$, the effective type-I error is already two times higher than the significance level $\alpha = 0.05$. For $\sigma_u^2 = 5$, the underestimation of the standard errors further increases. As expected, the bias does not improve with increased sample size.

Given the poor performance of the Poisson standard errors, the three alternative ways to calculate robust standard errors offer a clear improvement. For the larger sample, the nominal size of the test is closely realized even in the case of extreme overdispersion. The observed Poisson type-I error of 40.8% is reduced to 12.5% ($\widehat{\text{Var}}_{White}$), 10.9% ($\widehat{\text{Var}}_{lvf}$), and 11.4% ($\widehat{\text{Var}}_{qvf}$), respectively. Also in the small sample the performance of the three robust test procedures is much better than the one based on Poisson standard errors for $\sigma_u^2 = 1$ and $\sigma_u^2 = 5$ and is comparable to the performance of the test based on Poisson standard errors for $\sigma_u^2 = 0.2$.

These experiments demonstrate how misleading the use of Poisson standard errors can be in the presence of overdispersion, and how well the robust standard errors perform already in a medium sized sample. Not surprisingly, they also indicate a slight superiority of the t-test based on the assumption of a quadratic variance function (which is the correct one) which realizes closest the nominal size of the test in most, though not in all, of the experiments.

3.3.4 Non-Parametric Variance Estimation

Delgado and Kniesner (1997) propose an estimation method for a Poisson model with variance of unknown form that relies on generalized least squares using non-parametric estimation of the conditional variance. In particular, assume that y_i is distributed with mean $\mathrm{E}(y_i|x_i) = \exp(x_i'\beta)$ and conditional variance

$$\sigma_i^2 = \mathrm{Var}(y_i|x_i) \tag{3.46}$$

of unspecified functional form. The conditional variances can be estimated non-parametrically as

Table 3.2. Simulation Study for Poisson-PMLE: n=100

	$\sigma_u^2 = 0.2$	$\sigma_u^2 = 1$	$\sigma_u^2 = 5$	
Mean $\hat{\beta}_1$	1.0282	0.9900	1.0351	
Std. Deviation $\hat{\beta}_1$	0.4997	0.5606	0.9045	
	$\widehat{\text{Var}}_{\text{psn}}$	$\widehat{\text{Var}}_{\text{White}}$	$\widehat{\text{Var}}_{\text{lvf}}$	$\widehat{\text{Var}}_{\text{qvf}}$
$\sigma_u^2 = 0.2$				
t-test ($\alpha = 0.10$)	0.135	0.118	0.110	0.108
t-test ($\alpha = 0.05$)	0.058	0.066	0.049	0.056
$\sigma_u^2 = 1$				
t-test ($\alpha = 0.10$)	0.184	0.103	0.085	0.126
t-test ($\alpha = 0.05$)	0.105	0.055	0.044	0.064
$\sigma_u^2 = 5$				
t-test ($\alpha = 0.10$)	0.342	0.150	0.125	0.121
t-test ($\alpha = 0.05$)	0.272	0.087	0.069	0.064

Table 3.3. Simulation Study for Poisson-PMLE: n=1000

	$\sigma_u^2 = 0.2$	$\sigma_u^2 = 1$	$\sigma_u^2 = 5$	
Mean $\hat{\beta}_1$	0.9947	0.9953	0.9975	
Std. Deviation $\hat{\beta}_1$	0.1507	0.1754	0.2927	
	$\widehat{\text{Var}}_{\text{psn}}$	$\widehat{\text{Var}}_{\text{White}}$	$\widehat{\text{Var}}_{\text{lvf}}$	$\widehat{\text{Var}}_{\text{qvf}}$
$\sigma_u^2 = 0.2$				
t-test ($\alpha = 0.10$)	0.116	0.095	0.087	0.101
t-test ($\alpha = 0.05$)	0.057	0.043	0.045	0.048
$\sigma_u^2 = 1$				
t-test ($\alpha = 0.10$)	0.171	0.087	0.085	0.102
t-test ($\alpha = 0.05$)	0.103	0.045	0.039	0.048
$\sigma_u^2 = 5$				
t-test ($\alpha = 0.10$)	0.408	0.125	0.109	0.114
t-test ($\alpha = 0.05$)	0.323	0.065	0.055	0.058

$$\hat{\sigma}_i^2 = \sum_{j=1}^{k} (y_j - \exp(x_j' \tilde{\beta}))^2 w_{ij} \tag{3.47}$$

where $\tilde{\beta}$ is a root-n consistent estimator (for instance the Poisson PMLE), and w_{ij} are non-parametric k nearest neighbors probabilistic weights (see Delgado and Kniesner, 1997, for further details).

The semiparametric weighted least squares estimator β_0 is then obtained as a solution to

$$\sum_{i=1}^{n} \frac{(y_i - \exp(x_i'\beta))\exp(x_i'\beta)x_i}{\hat{\sigma}_i^2} = 0 \tag{3.48}$$

Delgado and Kniesner (1997) show that this estimator reaches the semiparametric efficiency bound, and state conditions for asymptotic normality. The model is applied to the number of work absence days in a sample of London Bus conductors and drivers during the early eighties. Robustness checks show that parameter estimates are sensitive to the choice of regressors but insensitive to the adopted econometric technique, including how the variance function is specified.

3.3.5 Poisson Regression and Log-Linear Models

Poisson pseudo maximum likelihood estimates the parameters of a correctly specified log-linear mean function consistently, even if higher order moment restrictions of the Poisson model do not hold. Most importantly, no assumptions on the variance function are required. Therefore, any process with mean function

$$\mathrm{E}(y|x) = \exp(x'\beta) \tag{3.49}$$

can be estimated consistently using the Poisson regression model. In particular, y does not need to be an integer, but it can be a non-negative continuous variable as well, and there may be great advantages of actually using the Poisson model in such instances, as pointed out by Santos Silva and Tenreyro (2006).

Traditionally, the recommendation has been to estimate multiplicative models with non-negative continuous dependent variables after taking logarithms. For example, if we write the regression model with multiplicative error as

$$y = \exp(x'\beta)\eta \tag{3.50}$$

then

$$\log y = x'\beta + \log \eta \tag{3.51}$$

A typical example is the Mincerian log-linear wage equation. There are two problems with this approach. First, it requires the dependent variable to be strictly positive. This may not be a problem, if the linearization is applied to wages of workers. In other applications, such as Cobb-Douglas type gravity models in trade, however, zero trade volumes between two countries are not unusual. While this problem is obvious, and can be dealt with in an ad-hoc manner, for example by adding a small constant, a second problem is more hidden but at least equally pernicious.

The mean function (3.49) implies that $\mathrm{E}(\eta|x) = 1$ in (3.50). But in the linearized version, $\mathrm{E}(\log \eta|x)$ is constant only if η is statistically independent of the regressors. If the variance (or higher order moments) of η depends on

x, the expected value of $\log \eta$ will also depend on the regressors. In this case, the linearized version (3.51) suffers from endogeneity, and estimation by OLS will provide inconsistent estimates of the semi-elasticities in (3.50).

To repeat the point, a necessary condition for OLS of the linearized model to be consistent is that η, and therefore $\log \eta$, is homoskedastic. A necessary and sufficient condition for consistency is that η, and therefore $\log \eta$, is statistically independent of x. Under independence, it follows that

$$\mathrm{Var}(y|x) = [\exp(x'\beta)]^2 \mathrm{Var}(\eta|x) \propto [\mathrm{E}(y|x)]^2$$

which may or may not be true. While the variance of a non-negative variable must go to zero as the mean passes to zero, there is no a-priori or theoretical reason to assume that $\mathrm{Var}(y|x)$ should be proportional to $[\mathrm{E}(y|x)]^2$.

In summary, it appears undesirable to use an estimator for the semi-elasticities β in (3.50) that so critically depends on the homoscedasticity assumption and is - unlike standard OLS estimation of the linear model - not robust to misspecification of the variance function. Santos Silva and Tenreyro (2006) therefore strongly recommend to estimate the multiplicative model directly.

There are a number of ways to proceed. Non-linear least squares is one, Poisson and gamma pseudo maximum likelihood estimator are two others. While all three estimators are consistent, they differ in how they weight the residuals. The Poisson estimator gives equal weight to all observations, while the non-linear least squares estimator gives more weight to observations with larger mean, which also tend to be observations with larger variance. As a consequence, the small sample behavior of these estimators differ. Santos Silva and Tenreyro (2006) show in a Monte Carlo simulation study that the Poisson PMLE seems to have the best properties overall. They conclude that "the Poisson PML estimator has the essential characteristics needed to make it the new workhorse for the estimation of constant-elasticity models." (p. 649).

3.3.6 Generalized Method of Moments

Our discussion in the preceeding sections has already shifted away from a fully parametric likelihood-based approach to an alternative one that is commonly referred to as *semi-parametric*, as parameters of interest are identified, and estimated, from a few moment conditions, without specifying the full data generating process. While pseudo maximum likelihood formally still involves the maximization of a likelihood function, the essence of it lies in the first order condition. The first order condition can be interpreted as a moment condition related to the correctly specified mean function $\mathrm{E}(y|x)$, and thus the orthogonality of the implied residuals $y - \mathrm{E}(y|x)$ and the regressors x.

It is not surprising then that alternative estimation methods can be, and have been, used in lieu of pseudo maximum likelihood. Non-linear least squares was mentioned earlier on, but the more general, encompassing estimation

framework for such semi-parametric models is the *generalized method of moments* (GMM). This method, the origins of which can be found in Hansen (1982), has become increasingly popular in count data modeling as well. While it sometimes offers only a re-interpretation of existing methods – in fact, the Poisson PML estimator is identical to the GMM estimator under some circumstances, see below – its genuine appeal comes from its potential to deal with non-standard sampling conditions, such as those related to endogeneity (Windmeijer and Santos Silva, 1997) or dynamic panel data modeling (Montalvo, 1997), that is, in the area of panel data models with weakly exogenous regressors (Chap. 7.2.5). Moreover, GMM provides a natural framework in which to conduct specification tests by way of testing for overidentifying restrictions (See, for instance, Santos Silva and Windmeijer, 1999).

Let θ be a $(p \times 1)$ vector of parameters that is to be estimated, and assume that there are l moment restrictions

$$\mathrm{E}[m_i(y_i, x_i; \theta)] = 0 \tag{3.52}$$

A well-known example for a set of moment restrictions is the instrumental variable estimator for the linear regression model, where it is postulated that $\mathrm{E}(z_i u_i) = 0$, where $u_i = y_i - x_i'\beta$ and z_i is a vector of instruments. The dimension of z_i may exceed that of the number of parameters (over-identification), or it may be just equal to the number of parameters, in which case the model is just identified.

The GMM estimator $\hat{\theta}$ minimizes the quadratic form

$$m(y, x; \theta)' A_n m(y, x; \theta) \tag{3.53}$$

where

$$m(y, x; \theta) = \sum_{i=1}^{n} m_i(y_i, x_i; \theta) \tag{3.54}$$

In the panel case, the simple sum over i in (3.54) has to be replaced by a double sum over both i and t. m is an $(l \times 1)$ vector of empirical moment restrictions and A_n is an $(l \times l)$ positive definite symmetric weighting matrix such that $\lim_{n\to\infty} A_n = A$. For $l = p$, the model is just identified and the empirical moment conditions can be solved directly so that the objective function at $\hat{\theta}$ will have a value of zero. However, the real advantage of GMM arises in situations of overidentification where $l > p$, i.e., the number of moment conditions exceeds the number of parameters. In this case, the various moment conditions will usually be conflicting and the minimization of (3.53), for a particular choice of weighting matrix A_n to be discussed below, combines the information provided by the different moment conditions in an optimal way.

In order to differentiate the objective function with respect to θ, note that

$$\frac{\partial m' A_n m}{\partial \theta} = 2 \left[\frac{\partial m'}{\partial \theta} \right] A_n m$$

Thus, the p first order conditions of the GMM estimator can be written as

$$D' A_n m = 0 \tag{3.55}$$

where

$$D = \sum_{i=1}^{n} \frac{\partial m_i(y_i, x_i; \theta)}{\partial \theta'}$$

is a $(l \times p)$ matrix. Numerical methods are usually required in order to solve the first-order conditions. If there is a unique vector θ satisfying

$$\mathrm{E}[m_i(y_i, x_i; \theta)] = 0$$

the model is identified, and $\hat{\theta}$ is the GMM estimator of θ.

Under mild regularity conditions, $\hat{\theta}$ is consistent and normally distributed with asymptotic covariance matrix (see, for instance, Davidson and McKinnon, 1993, Chap. 17)

$$\mathrm{Cov}(\hat{\theta}) = (D'AD)^{-1} D'AWAD(D'AD)^{-1}$$

where W is the covariance matrix of the specified moment restrictions

$$W = \mathrm{E}[m(y_i, x_i; \theta) m(y_i, x_i; \theta)']$$

If $A = W^{-1}$ the estimator is asymptotically efficient in the class of the given moment restrictions, and the asymptotic covariance matrix simplifies to

$$\mathrm{Cov}(\hat{\theta}) = (D'W^{-1}D)^{-1}$$

W in general depends on θ. In order to obtain a consistent estimator for $\hat{W}$, one can for instance obtain a consistent estimate $\hat{\theta}$ using any positive definite weighting matrix such as the identity matrix. When observations are independent over i the covariance matrix $\hat{W}$ can be calculated as

$$\hat{W} = \frac{1}{n} \sum_{i=1}^{n} m_i(\hat{\theta}) m_i(\hat{\theta})'$$

In a next step, the objective function $m'\hat{W}^{-1}m$ is minimized in order to obtain $\hat{\theta}$. The covariance matrix can be consistently estimated by $\widehat{\mathrm{Cov}}(\hat{\theta}) = D'\hat{W}^{-1}D$. Rather than doing just two steps, the process can be iterated to convergence (which can make a big difference in finite samples).

It was mentioned before that one of the advantages of GMM is that it offers a simple specification test for the validity of the overidentifying restrictions. If the model that led to the moment equations is incorrect, at least some of the sample moment conditions will by systematically violated. This provides the basis for a test. The statistic is simply the minimum of the criterion function evaluated at $\theta = \hat{\theta}$ and divided by the sample size n (division by n is not required if m is defined as the sample mean of the empirical moment conditions rather than the sample sum as in (3.54)). Hansen (1982) showed that under the null hypothesis $H_0 : \mathrm{E}[m_i(y_i, x_i; \theta)] = 0$, this test statistic has

a chi-squared distribution with degrees of freedom equal to $l - p$. This test is clearly non-directional as a rejection of the null hypothesis could be due to any of the moment conditions, ranging from failure of higher order moment conditions (if employed) to misspecified mean functions.

Application to the Poisson Model

Consider the key assumption of the Poisson regression model, $\lambda = \exp(x'\beta)$. Thus $\mathrm{E}(y|x) = \exp(x'\beta)$ and

$$y = \exp(x'\beta) + u$$

where $\mathrm{E}(u|x) = \mathrm{E}(y - \exp(x'\beta)|x) = 0$. Mean independence implies zero correlation, so that we obtain the orthogonality condition

$$\mathrm{E}[(y - \exp(x'\beta))x] = 0 \tag{3.56}$$

If x is of dimension $k \times 1$, (3.56) gives us k moment conditions, and these are sufficient to just identify the k regression parameters of the model, provided the regressors satisfy the usual full-rank condition. Mean independence is actually a much stronger assumption, as it implies that u is uncorrelated with *any* function of the covariates, for example squares x_k^2 or interactions $x_l x_k$.

The GMM estimator based on (3.56) yields the objective function

$$\left(\sum_{i=1}^{n}(y_i - \exp(x_i'\beta))x_i'\right) A \left(\sum_{i=1}^{n}(y_i - \exp(x_i'\beta))x_i\right) \tag{3.57}$$

where A is a weight matrix. As the minimum of (3.57) in this just identified model is obtained for

$$\sum_{i=1}^{n}(y_i - \exp(x_i'\beta))x_i = 0$$

the GMM estimator is identical to the Poisson maximum likelihood estimator, independently of the weighting matrix. With optimal weighting matrix equal to the inverse of the expectation of the outer product of the moment condition (the expected outer product of the score in the Poisson ML interpretation)

$$A = W^{-1} = \left(\sum_{i=1}^{n}\mathrm{Var}(y_i|x_i)x_i x_i'\right)^{-1}$$

and

$$D = \sum_{i=1}^{n}\frac{\partial(y_i - \exp(x_i'\beta))x_i}{\partial\beta'} = -\sum_{i=1}^{n}\exp(x_i'\beta)x_i x_i'$$

(D is the Hessian in the Poisson ML interpretation) the covariance matrix of the GMM estimator is

$$\mathrm{Cov}(\hat{\beta}_{GMM}) = \left[\sum_{i=1}^{n} \exp(x_i'\beta) x_i x_i' \left(\sum_{i=1}^{n} \mathrm{Var}(y_i|x_i) x_i x_i'\right)^{-1} \sum_{i=1}^{n} \exp(x_i'\beta) x_i x_i'\right]^{-1}$$

The variance of the GMM estimator is thus identical to the variance of the Poisson pseudo likelihood estimator, the usual sandwich formula. If the variance is equal to the mean, the variances of both are equal to the variance of the Poisson maximum likelihood estimator.

3.4 Sources of Misspecification

In Chap. 3.1.2 the specification of the Poisson regression model was introduced in three steps: the distributional assumption, the regression (mean function), and the assumption of independent sampling. A *misspecification* is a violation of any of the three assumptions. They can be dealt with in this order, although specific types of misspecification may affect not only one assumption, but two or three at a time.

The close relation between regression, variance function, and distribution is a particular feature of the Poisson regression model. Under the Poisson assumption, the equality of conditional mean and variance implies the loss of one degree of freedom as compared to, for instance, the normal linear model. Thus, a violation of the variance function always implies a violation of the distributional assumption. The violation that has obtained most attention is *overdispersion*, a situation where the variance exceeds the mean, conditional on covariates (see, among others, Cameron and Trivedi, 1990, Dean and Lawless, 1989, and Ganio and Shafer, 1992).

3.4.1 Mean Function

Recall that the mean function of the Poisson regression is specified as

$$\mathrm{E}(y|x) = \lambda = \exp(x'\beta) \tag{3.58}$$

where x is a vector of individual covariates. The main benefits of such an exponential mean function are threefold: it

- automatically respects the (non-negative) range of the dependent variable,
- provides an easy interpretation of coefficients in terms of semi-elasticities,
- leads to computationally simple expressions for the log-likelihood and its derivatives.

Nevertheless, there is clearly no law of nature telling us that the mean function must be log-linear, and potential sources of misspecification are manifold:

- The mean function is non-linear in β.
- Explanatory variables enter the mean function via some transformation $f(x)$, rather than linearly.

- The link function is misspecified. For instance, the true mean function may be linear rather than log-linear.

In practice, there can be little hope that the mean function is correctly specified in all its aspects, except the most trivial cases, for instance when the only regressor is a binary indicator variable. One approach to deal with this situation is to view the mean function (3.58) as a log-linear approximation to the true underlying mean function. Another approach is to explore more general functional forms.

An example along those lines, discussed by Wooldridge (1992) and Kenkel and Terza (2001), uses the inverse Box-Cox transformation. It introduces one additional parameter $\omega \in \mathbb{R}$, such that

$$\mathrm{E}(y|x) = [1 + \omega(x'\beta)]^{1/\omega}$$

This specification nests the linear model ($\omega = 1$) and the exponential model ($\omega = 0$).

Another way to generalize the functional form of the mean function is including higher order polynomials of all the regressors. Such a polynomial model can also be used for testing the functional form. However, with many regressors, a better approach is to include powers of the linear predictor, $(x'\hat{\beta})^2$, and $(x'\hat{\beta})^3$, say, in an auxiliary Poisson regression in analogy to the Reset test (Sapra, 2005).

Finally, methods for estimating generalized additive Poisson models are discussed in Hastie and Tibshirani (1986). Generalized additive models are very flexible, and can provide an excellent fit in the presence of nonlinear relationships. On the downside, there is a danger of "over-fitting" the data, i.e., obtaining results that likely cannot be replicated in alternative samples. In addition, such models lack the straightforward interpretation of a generalized linear model, and the results are therefore harder to understand and to communicate to others.

3.4.2 Unobserved Heterogeneity

Unobserved heterogeneity is an issue in count data modeling because the standard Poisson regression model makes no allowance for it. The rate, at which events occur,

$$\lambda = \exp(x'\beta)\,,$$

is a deterministic function of the regressors. The dependent variable is random, conditional on λ because $x'\beta$ determines only the rate at which events occur, but not the event counts themselves, which are subject to the intrinsic randomness of the Poisson process. If there are other variables that affect the rate at which events occur but are unobserved by the econometrician, and thus unaccounted for in the specification of the rate, we face a problem of unobserved heterogeneity.

Since we assume that observed regressors enter multiplicatively, it appears reasonable to make the same assumption for unobservables, and let

$$\tilde{\lambda} = \exp(x'\beta + v),$$

We might as well write

$$\tilde{\lambda} = \exp(x'\beta)u$$

where $u = \exp(v)$, our canonical form from now on. Together with the Poisson assumption, we obtain the modified conditional distribution model

$$f(y|x,u) = \frac{e^{-\lambda u}(\lambda u)^y}{y!}$$

which depends now on u in addition to x. Note that we could have formulated instead an additive model $\tilde{\lambda} = \exp(x'\beta) + \varepsilon$. But the additive approach is awkward to work with for at least two reasons. First, it treats observed and unobserved regressors asymmetrically, and there is no a-priori reason, why this should be so. Second, it imposes the restriction $\varepsilon > -\exp(x'\beta)$ which is inconvenient as well.

If u is independently distributed of x, with mean 1 (or any other constant – this restriction leads to no loss of generality as long as x includes a constant term) and variance σ_u^2, then we can obtain the mean and variance of y unconditional on u as

$$\mathrm{E}(y|x) = \mathrm{E}_u\mathrm{E}(y|x,u) = \lambda$$

and, an application of the variance decomposition theorem,

$$\mathrm{Var}(y|x) = \mathrm{E}_u\mathrm{Var}(y|x,u) + \mathrm{Var}_u\mathrm{E}(y|x,u) = \lambda + \sigma^2\lambda^2$$

Thus, $\mathrm{Var}(y|x) > \mathrm{E}(y|x)$, and we obtain the result that unobserved heterogeneity implies overdispersion.

By definition, u and thus $\tilde{\lambda}$ are unobserved. Assume, however, that we at least know the distribution of u, i.e., its density function $g(u)$. As $\tilde{\lambda}$ must be non-negative to qualify as a mean parameter of a count data distribution, we require that the support of $g(u)$ be the positive real numbers. By applying the basic change of variable technique where $\tilde{\lambda} = r(u) = \lambda u$ and $u = f(\tilde{\lambda}) = \tilde{\lambda}/\lambda$, we obtain that

$$\begin{aligned} h(\tilde{\lambda}) &= g[f(\tilde{\lambda})] \left| \frac{\mathrm{d}f(\tilde{\lambda})}{\mathrm{d}\tilde{\lambda}} \right| \\ &= g(\tilde{\lambda}/\lambda)/\lambda \end{aligned}$$

It is an arbitrary choice whether unobserved heterogeneity is introduced directly via $\tilde{\lambda}$ or indirectly via u. The latter option is slightly more common. If the marginal distribution $g(u)$ is given, we can express the joint density of y and u as

$$f(y,u) = f(y|u)g(u) \tag{3.59}$$

Finally, the marginal distribution for y is obtained by integrating the joint distribution $f(y, u)$ over u:

$$f(y) = \int_0^\infty f(y, u)\, du = \int_0^\infty f(y, \tilde{\lambda})\, d\tilde{\lambda} \tag{3.60}$$

For instance, if $f(y|u)$ is of the Poisson form, we get that

$$f(y) = \frac{\lambda^y}{y!} \int e^{-\lambda u} u^y g(u)\, du \tag{3.61}$$

Parametric models for unobserved heterogeneity specify the density function $g(u)$ is specified up to some unknown parameter(s). The leading examples are discussed in Chapter 4.

Spell-Specific Heterogeneity

Gourieroux and Visser (1997) have introduced the concept of *spell-specific heterogeneity*, modeling the counts as outcome of an underlying sequence of exponentially distributed spells (waiting times until the next occurrence). They use the fact that the probability that at most $k-1$ events occurred in a given interval $(0, T)$ equals the probability that the arrival time of the k-th event, given by the sum of the k waiting times τ_k between consecutive events, exceeds T. Moreover, assume that the waiting times τ_k $k = 1, 2, \ldots$ follow independent exponential distribution functions with parameters

$$\tilde{\lambda} = \tilde{\lambda}(x, u, \eta_k) \tag{3.62}$$

In addition to the two individual specific factors, the observed (x) and the unobserved (u), an additional spell-specific (unobserved) heterogeneity factor η_k is introduced. Gourieroux and Visser (1997) show that the underlying count data distribution, derived from a convolution operation and a local approximation of the characteristic function, can display both under- and overdispersion.

3.4.3 Measurement Error

Let $y|x$ be Poisson distributed with mean $\exp(x'\beta)$. One possible way of introducing measurement error in explanatory variables is to assume that rather than observing x, we observe

$$z = x + \varepsilon$$

where ε are assumed to be independent of x with mean 0 and covariance matrix Ω. Guo and Li (2001, 2002) study the consequences of such a set-up, and possible remedies. First, they note that measurement error leads to overdispersion for the observed model $f(y|z)$, much as unobserved heterogeneity does. We can write

$$f(y|z) = \int f(y|x)g(x|z)\mathrm{d}x$$

Unless $g(x|z) = 0$ almost everywhere (or $\mathrm{E}(y|z) = 1$), it can be shown that

$$\mathrm{E}(y|z) < \mathrm{Var}(y|z)$$

even though $\mathrm{E}(y|x) = \mathrm{Var}(y|x)$. Clearly, thus, the Poisson model is inappropriate. However, in contrast to standard unobserved heterogeneity, the problem goes beyond considerations of efficiency and consistent estimation of the covariance matrix for valid inference. Rather, the Poisson estimator $\hat{\beta}$ is inconsistent in general. To see the thrust of the argument, re-write the log-likelihood function of the falsely assumed Poisson model $f(y|z)$

$$\sum_{i=1}^{n}[-\exp(z_i'\beta) + y_i z_i'\beta - \log(y_i!)]$$

using $z_i = x_i + \varepsilon_i$, as

$$\sum_{i=1}^{n}[-\exp(x_i'\beta) + y_i x_i'\beta - \log(y_i!)]$$

$$+\sum_{i=1}^{n}[-\exp(x_i'\beta)[\exp(\varepsilon_i\beta) - 1] + y_i\varepsilon_i\beta]$$

Consistent parameter estimation could be based on the first (unobserved) term of the log-likelihood function. Since the likelihood function of the Poisson model with measurement error adds a second term, its maximization in general will not yield a consistent estimator. Note that the second term converges to

$$n\mathrm{E}_x[-\exp(x'\beta)](\mathrm{E}_\varepsilon[\exp(\varepsilon\beta)] - 1) = n\mathrm{E}_z[-\exp(z'\beta)] + n\mathrm{E}_x[\exp(x'\beta)]$$

This suggests that to obtain a consistent estimator, one can possibly maximize

$$\frac{1}{n}\sum_{i=1}^{n}[-\exp(z_i'\beta) + y_i z_i'\beta - \log(y_i!)] - \frac{1}{n}\sum_{i=1}^{n}[-\exp(z_i'\beta)] - \mathrm{E}_x[\exp(x'\beta)]$$

$$= \frac{1}{n}\sum_{i=1}^{n}[y_i z_i'\beta - \log(y_i!)] - \mathrm{E}_x[\exp(x'\beta)]$$

given that $\mathrm{E}_x[\exp(x'\beta)]$ is known or can be estimated first. Guo and Li (2002) refer to the estimator that maximizes this modified log-likelihood function by solving

$$\frac{1}{n}\sum_{i=1}^{n} y_i z_i - \mathrm{E}_x[x\exp(x'\beta)] \equiv 0$$

as *corrected score estimator*. The crucial question is now how to obtain $\mathrm{E}_x[\exp(x'\beta)]$. For example, when ε is multivariate normal with mean 0 and covariance Ω, then

$$\mathrm{E}_x[\exp(x'\beta)] = \mathrm{E}_z\left[\exp\left(z'\beta - \frac{\beta'\Omega\beta}{2}\right)\right]$$

which can be consistently estimated by

$$1/n\sum_{i=1}^{n}\exp\left(z_i'\beta - \frac{\beta'\Omega\beta}{2}\right) ,$$

so that the corrected log likelihood function can be written as

$$\sum_{i=1}^{n}\left[y_i z_i'\beta - \log(y_i!) - \exp\left(z_i'\beta - \frac{\beta'\Omega\beta}{2}\right)\right] \tag{3.63}$$

See Guo and Li (2002) and Nakamura (1990) for further details on this approach. Guo and Li (2001) study measurement error in the negative binomial model. In contrast to the classical measurement error model, the direction of the bias is not necessarily one of attenuation. The methods can be extended to deal with proxy variables rather then measurement error. Finally, Cameron and Trivedi (1998) discuss the case of multiplicative measurement error in explanatory variables.

3.4.4 Dependent Process

The Poisson distribution is appropriate if events occur completely randomly over time. The Poisson model is misspecified if the process that generates events is not "memoryless", that is, when the probability of an occurrence between t and $t+h$ depends on the past process. Few attempts have been made in the literature to directly model the dependence structure. One such model is a "contagious" process that leads to the negative binomial distribution. Another is the Winkelmann (1995) model for duration dependence in count processes. Starting point is the observation that the Poisson distribution is characterized by independently exponentially distributed interarrival times of events. The key feature of the exponential distribution is its constant hazard function, i.e., lack of memory. Winkelmann (1995) shows how a count data distribution can be derived based on a more general distribution of interarrival times, the gamma distribution. In particular, the gamma distribution allows for both negative duration dependence, leading to overdispersion, and positive duration dependence, leading to underdispersion.

3.4.5 Selectivity

While there is a vast literature on sample selection in the linear model, the analysis of count data models with selectivity is less well developed. Sample selection occurs if the data are generated such that the researcher does not observe the underlying count variable y^* but rather a "selected" count y. The types of selection considered include truncation and censoring, but also

underreporting (see Chap. 3.4.7). If sample selection is ignored the estimator of the regression parameters is generally inconsistent.

Two patterns of selectivity can be distinguished. Firstly, observations can be censored or truncated depending on the outcome of y^*. For instance, many survey questionnaires are "top-coded", introducing a category of the type "x or more" events. In this case, the data are censored from above. Secondly, observations can be censored or truncated depending on the outcome of another variable c that may or may not be independent of y^*. The literature refers to this case as "incidental" truncation or censoring (Greene, 2000). We denote this situation as "endogenous" selectivity. An example is the study of credit card defaults. Incidental truncation occurs since some individuals have no access to credit cards. Some information on the joint distribution of y^* and c is required in this situation. Such models are discussed, among others, in Terza (1998), and Winkelmann (1998).

3.4.6 Simultaneity and Endogeneity

The basic Poisson regression model was introduced as a single equation regression model for cross section data. In a next step, unobserved heterogeneity was allowed for. In contrast to the linear model where unobserved heterogeneity is automatically taken into account, we showed that the Poisson model was inappropriate in this situation and had to be generalized. The generalization was based on the key assumption that unobserved heterogeneity and regressors x were statistically independent.

This assumption is likely to be violated in many applications, in particular, when regressors are simultaneously determined and hence endogenous. The prime example for endogenous regressors is an endogenous treatment effect, where individuals self-select into treatment, and those who take the treatment are systematically different from the control group. If this selection is correlated with the outcome, either directly or through unobservables, the assumption of statistical independence between regressors and the error term will break down and standard estimation methods like maximum likelihood will not generally be consistent. The generic solution to this problem is a nonlinear instrumental variable approach as outlined in Mullahy (1997a) and in Windmeijer and Santos Silva (1997). Alternatively, one may attempt to directly model the endogenous regressor and employ two-stage estimation techniques. This is described in Terza (1998).

A problem of a different sort is the modeling of multivariate counts. For instance, Chap. 7.1.1 introduces a bivariate Poisson model. While this model does not give rise to a simultaneous system (since none of the dependent variables appears as a regressor), it constitutes what has come to be known in the literature on linear models as the seemingly unrelated regression model. In particular, the bivariate Poisson model allows for a non-trivial correlation structure between the two or more endogenous count variables. If correlation

is present, joint estimation will lead to a more efficient estimator than separate estimation.

3.4.7 Underreporting

Although the issue of underreporting can be interpreted in terms of selectivity, it leads to models of a different type. Count data are underreported if only a fraction of the total events is reported. If y^* denotes the total number of events, and y the number of reported events, then clearly $y \leq y^*$. In a different interpretation, y^* denotes the number of potential events and y the number of actual events. For instance, in labor economics, y^* could be the number of job offers during a given period of time, in which case y as the number of accepted offers and $y^* - y$ is the number of rejected offers. Both interpretations have the same formal structure. In particular,

$$y = \sum_{j=1}^{y^*} B_j \tag{3.64}$$

where B_i is an indicator variable taking the value 1 if an event is reported (or a job is accepted) and zero otherwise. In statistical terms, the distribution of y is referred to as a convolution (or stopped-sum distribution; see Chap. 2.5.2). The distribution of y depends on the joint distribution for y^* and the B_j's. In general, closed form results are only available under strong independence assumptions.

Three generic types of underreporting have been discussed in the literature, each giving rise to a different count data model:

1. Random underreporting. Here, the B_j's are independently and identically Bernoulli distributed with parameter p (Winkelmann, 1996).
2. Logistic underreporting. The probability of reporting $P(B_j = 1)$ is a logistic function of covariates (Winkelmann and Zimmermann, 1993, Mukhopadhyay and Trivedi, 1995).
3. Count amount model. Events are associated with a nonnegative size variable (for instance a purchase amount), and recorded only if this variable exceeds a specific minimum threshold (Van Praag and Vermeulen, 1993).

3.4.8 Excess Zeros

The idea of adjusting the probability of zero outcomes for count distributions goes at least back to Johnson and Kotz (1969). See also Mullahy (1986). An early application in a regression framework is Lambert (1992). She introduces a zero-inflated Poisson model where with probability ω the only possible observation is 0, and with probability $1 - \omega$ a Poisson(λ) random variable is observed. Both ω and λ may depend on covariates. The overall probability of a zero outcome is then

$$f(0) = \omega + (1-\omega)e^{-\lambda} \tag{3.65}$$
$$= e^{-\lambda} + \omega(1-e^{-\lambda})$$

This probability is strictly greater than the Poisson probability of 0 as long as $\omega > 0$. Excess zeros, like unobserved heterogeneity and dependence, lead to overdispersion. Hence, excess zeros provide yet another explanation for this frequently observed phenomenon. Moreover, the model has an interesting interpretation. Lambert (1992) studies the number of defects in manufacturing. Here, the count generating process is decomposed into a "perfect state" where defects are extremely rare and an "imperfect state" where defects are possible but not inevitable.

An economic interpretation is given in Crépon and Duguet (1997b) in a study of R&D productivity. Assume that the count variable is the number of patent applications lodged by a firm during a given period of time. Here, firms face the strategic choice whether or not to apply for patents in general. Only when this choice is made in the affirmative is it that the number of actual discoveries becomes relevant. Again, no applications may result for firms that decided to patent but had no discoveries during a particular period.

3.4.9 Variance Function

It was noted before that a variance violation implies a distributional violation. The opposite does not follow, since the distributional violation might originate in higher order moments. Here, the possibility of such higher order violations is left unexplored and the focus is on the variance function. A rationale for this approach is that most properties of the model and of the estimator are established through asymptotic results which require assumptions on the first two moments only.

The variance function of the benchmark Poisson model is $\mathrm{Var}(y|x) = \mathrm{E}(y|x) = \exp(x'\beta)$. Count data of this sort are said to be equi-dispersed. The two alternatives are *overdispersion* or *underdispersion*. In the former situation, the conditional variance exceeds the conditional mean; in the latter, the conditional mean exceeds the conditional variance. The following causes for non-Poissoness of the variance function can be distinguished:

- Unobserved individual heterogeneity causes overdispersion. This case has been discussed in Chap. 3.4.2.
- Spell specific heterogeneity as defined by Gourieroux and Visser (1997) may result in either over- or underdispersion.
- True positive contagion causes overdispersion; true negative contagion causes underdispersion.
- Non-stationarity has an ambiguous effect. If non-stationarity can be modeled as a convolution of independent Poisson distributions, the convolution is again Poisson distributed.

Over- and underdispersion exist if the function mapping the conditional mean into the conditional variance is not the identity function. In general, this may be an arbitrary function, possibly involving further individual attributes z:

$$\mathrm{Var}(y|x,z) = f[\mathrm{E}(y|x), z] \tag{3.66}$$

or, assuming that the mean function is correctly specified

$$\mathrm{Var}(y|x,z) = f[\exp(x'\beta), z] \tag{3.67}$$

The quadratic variance function

$$\mathrm{Var}(y|x) = \exp(x'\beta) + \sigma^2[\exp(x'\beta)]^2 \qquad \sigma^2 \in I\!R^+ \tag{3.68}$$

has received most attention in the literature. It arises naturally if the model has unobserved heterogeneity with constant variance.

Generalizations have proceeded in two directions. First, the range of σ^2 may include negative values, allowing for underdispersion (King, 1989b). Some constraints on the parameter space are required since the left side, a variance, has to be kept positive. Second, σ^2 may be parameterized in terms of explanatory variables. These variables usually coincide with those appearing in the mean function, although this is not a formal requirement. In count data models where mean and variance are intrinsically linked it would be difficult, however, to justify that some variables z affect the variance but not the mean. Thus, one common parameterization is

$$\tilde{\sigma}^2(x) = \sigma^2 \exp(x'\gamma) \tag{3.69}$$

The potential problem with this approach is that the model tends to be overparameterized. Winkelmann and Zimmermann (1991a) avoid this problem by imposing proportionality: $\gamma = (k-1)\beta$, where $(k-1)$ is a proportionality factor. Then

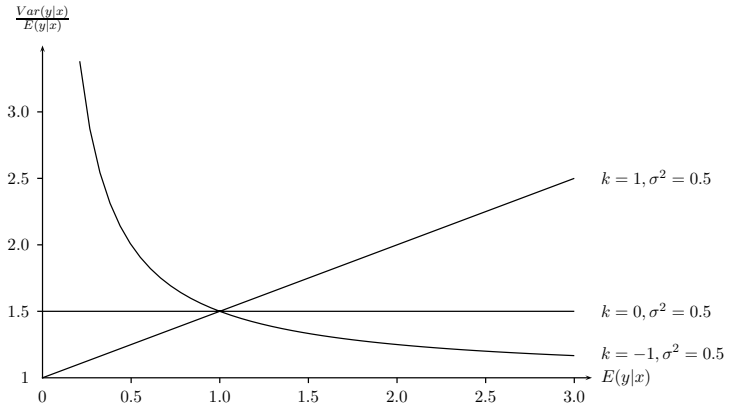
$$\begin{aligned}\tilde{\sigma}^2(x) &= \sigma^2 \exp[(k-1)x'\beta] \\ &= \sigma^2[\exp(x'\beta)]^{k-1}\end{aligned} \tag{3.70}$$

Therefore,

$$\mathrm{Var}(y|x) = \mathrm{E}(y|x) + \sigma^2[\mathrm{E}(y|x)]^{k+1} \qquad \sigma^2 \in I\!R^+,\ k \in I\!R\,. \tag{3.71}$$

σ^2 is the *dispersion*-parameter and k the non-linearity parameter. Compared to the Poisson variance assumption, this specification has two additional parameters. It allows for linear, convex as well as concave relationships between the variance and the mean. For $k = 0$ and positive σ^2, this is the variance function of the Negbin I model. For $k = 1$ and positive σ^2, this is the variance function of the Negbin II model. Fig. 3.3 shows some possible shapes of the variance function.

Fig. 3.3. Variance-Mean Relationships for Different k's and σ^2's



3.5 Testing for Misspecification

Since many of the misspecifications listed in the previous section lead to a violation of the assumption of equal conditional variance and mean, the variance function provides a natural starting point for misspecification tests. Frequently one has a specific alternative in mind that allows for a more general variance function and, at the same time, nests the Poisson variance function through a parametric restriction. In this situation, the three classical tests, the likelihood ratio, the Wald, and the Lagrange multiplier test can be used.

A related class of 'misspecification' tests addresses the inclusion of irrelevant variables in the mean function. The difference is that these restrictions are tested within a given parametric distribution family that is specified up to a parameter vector β which may contain zeros. As a consequence, these tests can rely on standard maximum likelihood technology and do not need a special treatment.

Tests that discriminate between non-nested models have a shorter history in econometrics. And yet, non-nested models frequently arise in applied work with count data. Examples include the hurdle Poisson and the zero-inflated Poisson models, or Poisson models with linear and log-linear mean functions. A simple asymptotic test for non-nested models has been introduced by Vuong (1989). See also Santos Silva (2001). A simulation based approach offers an alternative to discriminate between non-nested models. Finally, the Hausman (1978) test can be applied in some specific situations.

3.5.1 Classical Specification Tests

Testing a Poisson model against a more general parametric model is straightforward if the former is contained in the latter through a parametric restriction. The two models are said to be "nested". Examples for restrictions are:

1. Poisson versus negative binomial: $H_0 : \sigma_u^2 = 0$.
2. Linear restrictions on regression coefficients: $H_0 : R\beta = q$.
3. Non-linear restrictions on regression coefficients: $H_0 : R(\beta) = q$.

Assume that estimation is by the method of maximum likelihood. Tests for the validity of H_0 can be based on any one of the three following principles:

1. Likelihood-ratio test
2. Wald test
3. Lagrange multiplier test

These are asymptotic tests. Their small sample properties (size and power) are generally unknown. Asymptotically, all three tests are equivalent. The tests are directional, implying that if the null hypothesis is rejected, there is a well defined alternative.

Likelihood Ratio Test

Let $\hat{\ell}_r$ denote the value of the log-likelihood function evaluated at the restricted maximum likelihood estimates (for instance, the Poisson model), and $\hat{\ell}_u$ the value of the log-likelihood function evaluated at the unrestricted maximum likelihood estimates (for instance, the negative binomial model), and let k denote the number of restrictions (k=1 in a test of the Poisson model against the negative binomial model). Then, under H_0 (if the restriction is correct):

$$\text{LR} = -2(\hat{\ell}_r - \hat{\ell}_u) \sim \chi^2_{(k)} \tag{3.72}$$

where $\chi^2_{(k)}$ is a chi-squared distribution with k degrees of freedom. The test is based on the difference of two log-likelihood values, or, equivalently, the log of a likelihood-ratio, hence its name. If the restriction is lifted, the value of the log-likelihood function (evaluated at the maximum likelihood estimates) must increase. If the increase is "large", where large is defined as any test statistic that exceeds the critical value $\chi^2_{\alpha,k}$, the null hypothesis is rejected.

This test is sometimes criticized because it requires separate estimation of two models, the restricted model and the unrestricted model. However, given modern computing power, this criticism has lost some of its original weight.

One problem with the likelihood ratio test of the Poisson against the negative binomial model is that under the null hypothesis the true parameter is on the boundary of the parameter space. If a parameter, such as a variance, is bounded from below at H_0, the estimate must be greater than or equal to H_0 and vice versa. The asymptotic normality of the MLE can no longer hold under H_0. Chernoff (1954) and Lawless (1987b) address this problem. Chernoff (1954) shows that under H_0, the likelihood ratio statistic has a distribution with probability mass of 0.5 at 0 and a $0.5\chi^2_{(1)}$ distribution for positive values. This adjustment is not required when testing restrictions on regression coefficients.

Wald Test

Point of departure for the Wald test is the asymptotic distribution of the maximum likelihood estimator in the unrestricted model. In contrast to the likelihood-ratio test, estimation of a single model is sufficient:

$$\hat{\theta} \overset{\text{app}}{\sim} N(\theta_0, \widehat{\text{Var}}(\hat{\theta}))$$

where $\hat{\theta}$ consists of the estimated regression coefficients plus any additional parameters, such as $\hat{\sigma}_u^2$. For any linear combination of the parameter vector $R\hat{\theta} - q$ it follows that

$$R\hat{\theta} - q \overset{\text{app}}{\sim} N(R\theta_0 - q, R\widehat{\text{Var}}(\hat{\theta})R')$$

Under $H_0 : R\theta_0 - q = 0$ and therefore

$$\text{W} = (R\hat{\theta} - q)'[R\widehat{\text{Var}}(\hat{\theta})R']^{-1}(R\hat{\theta} - q) \tag{3.73}$$

has a chi-squared distribution with degrees of freedom equal to the number of restrictions if the null hypothesis is correct. If the number of restrictions is one, the statistic simplifies to the squared "t"-statistic. Dividing W in (3.73) by the degrees of freedom produces an "F"-statistic. The quotation marks indicate that, under maximum likelihood estimation, asymptotic results apply, whereas the t-distribution and the F-distribution are finite sample distributions.

To give an example, assume that estimation of a negative binomial model produces an estimate $\hat{\sigma}^2$ with estimated asymptotic variance $\widehat{\text{Var}}(\hat{\sigma}^2)$. The Poisson model requires $\sigma^2 = 0$. Hence, the Wald test of H_0 : Poisson(λ) against H_1: negative binomial with mean λ and variance $\lambda + \sigma^2\lambda^2$ is based on the "t"-statistic $(\hat{\sigma}^2 - 0)/\sqrt{\widehat{\text{Var}}(\hat{\sigma}^2)}$. Again, the parameter is at the boundary of the parameter space, and hence inference should use a one-sided alternative, based on half a standard normal distribution $N(0, 1)$.

Finally, if the restriction is of a non-linear nature, we can substitute

$$\text{Var}[R(\hat{\theta}) - q] = \left[\frac{\partial R(\hat{\theta})}{\partial \hat{\theta}'}\right] \widehat{\text{Var}}(\hat{\theta}) \left[\frac{\partial R(\hat{\theta})}{\partial \hat{\theta}}\right]$$

for the variance (an application of the "delta" rule), and $R(\hat{\theta}) - q$ for $R\hat{\theta} - q$ in equation (3.73). Under H_0, the resulting statistic is again approximately chi-squared distributed with k degrees of freedom.

Lagrange Multiplier Test

Instead of computing both models und performing a likelihood ratio test, or computing the alternative model only and performing a Wald test, the Lagrange multiplier (LM) test avoids the computation of the alternative model altogether. This test is also known as "score" test. In this context, the score vector is the vector of first derivatives (or score) of the log-likelihood function with respect to the parameters.

Let $\log L_u$ be the log-likelihood function of the unrestricted model. Then $\hat{\theta}_u$ solves the first-order conditions

$$\left.\frac{\partial \log L_u}{\partial \theta}\right|_{\theta_u} = \left.\frac{\partial \ell_u}{\partial \theta}\right|_{\theta_u} = 0 \tag{3.74}$$

Alternatively, we could evaluate the score vector of the unrestricted model at the maximum likelihood estimator $\hat{\theta}_r$ obtained under the restricted model:

$$\left.\frac{\partial \ell_u}{\partial \theta}\right|_{\theta_r} \neq 0 \tag{3.75}$$

Unless the restriction is true, this expression differs from zero. Hence, the restriction is rejected if the score (3.75) is "far" from zero. If the null hypothesis is true (i.e., $\hat{\theta}_r = \theta_0$), it was shown in Chap. 3.2 that the score vector has an asymptotic normal distribution with mean zero and variance covariance matrix equal to the Fisher information matrix. Hence, a test can be based on the quadratic form

$$\text{LM} = \left(\frac{\partial \ell_u}{\partial \theta}\right)'_{\hat{\theta}_r} [I(\hat{\theta}_r)]^{-1} \left(\frac{\partial \ell_u}{\partial \theta}\right)_{\hat{\theta}_r} \tag{3.76}$$

which has a chi-squared distributed with degrees of freedom equal to the number of restrictions under the null hypothesis. (The variance of the score can be estimated along the common three ways, see Chap. 3.2; See Greene (2000) for an explanation why this test statistic is related to a Lagrange multiplier.).

As an example, one could use the score/LM test for testing a linear restriction on the regression coefficient. Let $\hat{\beta}_r$ be the maximum likelihood estimator obtained from estimating the Poisson regression with restriction imposed. Let $\hat{\Lambda}_r = \exp(X\hat{\beta}_r)$ be the $(n \times 1)$ vector of predicted conditional means in the restricted model. The score/LM statistic is given by

$$\text{LM} = (Y - \hat{\Lambda}_r)' X (X' \text{diag} \hat{\Lambda}_r X)^{-1} X' (Y - \hat{\Lambda}_r) \tag{3.77}$$

Under the null hypothesis, this statistic is chi-squared distributed with r degrees of freedom.

Usually, however, score/LM tests are used in a different context: rather than testing restrictions on regression coefficients, score/LM procedures have been developed to test certain aspects of the stochastic specification.

A leading example is Lee (1986) who derives a Lagrange multiplier for the Poisson regression model against the more general Katz models. The Katz family contains the negative binomial model as a special case, and we shall illustrate the derivation of the LM statistic for the Poisson against the Negbin II model. Recall that the probability distribution function is given by

$$f(y) = \frac{\Gamma(\alpha + y)}{\Gamma(\alpha)\Gamma(y+1)} \left(\frac{\alpha}{\alpha + \lambda}\right)^{\alpha} \left(\frac{\lambda}{\alpha + \lambda}\right)^{y}$$

The relevant restriction is $H_0 : \alpha^{-1} = 0$. The derivatives of the log density with respect to α and λ are given in Chap. 4.3. Unfortunately, brute force evaluation of $\partial\ell/\partial\alpha|_{\alpha^{-1}=0}$ does not work in this case, and a more elaborate derivation is needed (See also Cameron and Trivedi, 1998, Chap. 5).

Using the product representation of the gamma ratio, the probability distribution function for a single observation can be written as

$$\begin{aligned} f(y) &= \prod_{j=1}^{y} \frac{j+\alpha-1}{j(1+\alpha/\lambda)}(1+\lambda/\alpha)^{-\alpha} \\ &= f(0)\prod_{l=1}^{y} \frac{\alpha+(y-l)}{(1+\alpha/\lambda)(y-l+1)} \\ &= f(0)\prod_{l=1}^{y} \frac{\lambda+a\lambda(y-l)}{(1+a\lambda)(y-l+1)} \end{aligned}$$

where $f(0) = (1+\lambda/\alpha)^{-\alpha}$, $l = y-j+1$ and $a = \alpha^{-1}$. The derivative of the log-density with respect to a is given by

$$\frac{\partial\ell}{\partial a} = \left\{\sum_{l=1}^{y} \frac{\lambda(y-l)}{\lambda+a\lambda(y-l)} - \frac{\lambda}{1+a\lambda}\right\} + \frac{\partial \log f(0)}{\partial a} \tag{3.78}$$

Under $H_0 : a = 0$,

$$\begin{aligned} \left.\frac{\partial\ell}{\partial a}\right|_{a=0} &= \left\{\sum_{l=1}^{y}(y-l) - \lambda\right\} + \left.\frac{\partial \log f(0)}{\partial a}\right|_{a=0} \\ &= \sum_{j=1}^{y}(j-l) - \lambda y + \left.\frac{\partial \log f(0)}{\partial a}\right|_{a=0} \end{aligned}$$

The sum is simply $y(y-1)/2$. Since for any proper distribution the expected score is zero, we have that $\mathrm{E}(\partial \log f(y)/\partial a) = \mathrm{E}(\partial h(y)/\partial a + \partial \log f(0)/\partial a) = 0$, and thus $\mathrm{E}(\partial \log f(0)/\partial a) = \partial \log f(0)/\partial a = -\mathrm{E}(h(y))$. Putting things together,

$$\begin{aligned} \left.\frac{\partial\ell}{\partial a}\right|_{a=0} &= y(y-1)/2 - \lambda y - \mathrm{E}[y(y-1)/2 - \lambda y] \\ &= y(y-1)/2 - \lambda y - (\lambda^2/2 - \lambda^2) \\ &= \frac{1}{2}[(y-\lambda)^2 - y] \end{aligned}$$

For a random sample of size n, and with $\hat{\lambda} = \exp(x'\hat{\beta})$ where $\hat{\beta}$ is the estimated vector of regression coefficients under the null hypothesis (i.e., the Poisson maximum likelihood estimator), the sample score can be written as

$$\left.\frac{\partial\ell_n}{\partial a}\right|_{a=0} = \frac{1}{2}\sum_{i=1}^{n}(y_i - \hat{\lambda})^2 - y_i \tag{3.79}$$

In addition, a consistent estimator for the variance can be obtained using the information matrix equality as

$$\frac{1}{n}\mathrm{E}\left(\frac{\partial \ell_n}{\partial a}\right)^2_{\hat{\beta},a=0} = \sum_{i=1}^{n} \frac{1}{2\hat{\lambda}_i^2}$$

and the square root of the scalar test statistic is given by

$$\mathrm{LM} = \left[\sum_{i=1}^{n} \frac{1}{2\hat{\lambda}_i^2}\right]^{-1/2} \frac{1}{2}\sum_{i=1}^{n}(y_i - \hat{\lambda}_i)^2 - y_i \tag{3.80}$$

Under H_0 the score has an asymptotic standard normal distribution (since it is the square root of a chi-squared distribution with one degree of freedom) and a test for overdispersion is a one sided test with critical value z_α. Since the same statistic is obtained for the more general Katz family, it can also be used to test for underdispersion. Rejection of equidispersion against underdispersion requires a test statistic smaller than $-z_\alpha$.

Score tests have been developed in the literature on count data models for all sorts of null and alternative hypotheses. For instance, Gurmu (1991) derives a score test for overdispersion in the positive Poisson regression models, Gurmu and Trivedi (1992) consider overdispersion in truncated Poisson regression models, while van den Broek (1995) develops a score test for extra zeros in the Poisson model. As in Chap. 2.4.3 the zero altered Poisson model can be written as

$$P(y_i = 0) = \omega + (1-\omega)e^{-\lambda_i}$$

$$P(y_i = k) = (1-\omega)\frac{e^{-\lambda_i}\lambda_i^k}{k!} \qquad k = 1, 2, \ldots$$

This model collapses to the standard Poisson model for $\omega = 0$. Let $\hat{\lambda}_i = \exp(x_i'\hat{\beta})$ where $\hat{\beta}$ are the usual Poisson estimates. van den Broek shows that under $H_0 : \omega = 0$,

$$\mathrm{LM} = \frac{\left(\sum_{i=1}^{n}(\mathbf{1}(y_i = 0) - e^{-\hat{\lambda}_i})/e^{-\hat{\lambda}_i}\right)^2}{\left(\sum_{i=1}^{n}(1 - e^{-\hat{\lambda}_i})/e^{-\hat{\lambda}_i}\right) - n\bar{y}} \tag{3.81}$$

where $\mathbf{1}$ is the usual indicator function, has a chi-squared distribution with 1 degree of freedom.

Information Matrix Test

The information matrix test (White, 1982) is based on the sample analogue of the identity

$$\mathrm{E}\left(\frac{\partial \ell}{\partial \theta}\frac{\partial \ell}{\partial \theta'}\right) = -\mathrm{E}\left(\frac{\partial^2 \ell}{\partial \theta \partial \theta'}\right)$$

which can be rewritten as

$$\mathrm{E}\left(\frac{\partial \ell}{\partial \theta}\frac{\partial \ell}{\partial \theta'} + \frac{\partial^2 \ell}{\partial \theta \partial \theta'}\right) = 0$$

In the context of count data models with unobserved heterogeneity, it has been noted by Lee (1986) that a result originally due to Chesher (1984) on the potential equality between information matrix tests and score/Lagrange multiplier tests applies to the Poisson regression model as well. From Chap. 3.2, the sample analogues for the Hessian and the variance of the outer product of the score vector of the Poisson regression model are given by the two expressions $\sum_{i=1}^{n} \hat{\lambda}_i x_i x_i'$ and $\sum_{i=1}^{n} (y_i - \hat{\lambda}_i)^2 x_i x_i'$, respectively. The difference between the two matrices depends on elements $(y_i - \hat{\lambda}_i)^2 - \hat{\lambda}_i$. In particular, the sample matrix difference is, up to a factor of 1/2, equal to the sample score evaluated at the restricted value (3.79) if the information test precedure is applied to the intercept parameter β_0. This follows, since $\sum_{i=1}^{n} (y_i - \hat{\lambda}_i)^2 - \hat{\lambda}_i = \sum_{i=1}^{n} (y_i - \hat{\lambda}_i)^2 - y_i$. Hence the interpretation of the information matrix test as a test for neglected heterogeneity (See also Moffatt, 1997a).

3.5.2 Regression Based Tests

In the same way as a residual analysis in the linear model with normally distributed errors can reveal heteroskedasticity, the Poisson residuals may indicate a violation of equidispersion. The analysis may proceed either graphically, or through auxiliary regressions. Define $\hat{\mathrm{E}}(y_i) = \hat{\lambda}_i$ and $\widehat{\mathrm{Var}}(y_i) = (y_i - \hat{\lambda}_i)^2 = \hat{u}_i^2$. Plotting $\hat{\mathrm{E}}(y_i)$ against $\widehat{\mathrm{Var}}(y_i)$ should produce points scattered around the 45^o line. Alternatively the regression (See Cameron and Trivedi 1986, 1990)

$$\widehat{\mathrm{Var}}(y_i) = \theta \hat{\mathrm{E}}(y_i) + \nu_i \tag{3.82}$$

should yield an estimate $\hat{\theta}$ close to 1. The regression

$$\frac{\widehat{\mathrm{Var}}(y_i)}{\hat{\mathrm{E}}(y_i)} = \theta_1 + \theta_2 \hat{\mathrm{E}}(y_i) + \nu_i \tag{3.83}$$

should yield an $\hat{\theta}_1$ close to 1 and a $\hat{\theta}_2$ close to 0.

3.5.3 Goodness-of-Fit Tests

In contrast to continuous modeling, discrete data allow to calculate probabilities of single outcomes after the model has been estimated. In the domain of binary variables, this fact has been recognized for a long time, and a comparison between actual and predicted outcomes is usually contained in the available statistical software. Prediction tables have been criticized for being uninformative, since the fitted model can be outperformed by a naive

model predicting all outcomes to be equal to the most frequent outcome in the sample (See, for instance, Veall and Zimmermann, 1992). For count data models, however, the situation is more favorable, although most of the applied literature has ignored the possibility of using the predictions to evaluate the goodness-of-fit. Notable exceptions are Gilbert (1982) and Dionne and Vanasse (1992). See also Alvarez and Delgado (2002).

Gilbert (1982) considers the Poisson model and measures the goodness-of-fit by the proportion of correct in-sample predictions. He suggests to predict the count for individual i (with given attributes x_i) either by its modal value, or the integer nearest to its expected value. This procedure thus basically rests on the goodness-of-fit of the mean function.

A related procedure is based on the Pearson statistic

$$P = \sum_{i=1}^{n} \frac{(y_i - \hat{\lambda}_i)^2}{\hat{\lambda}_i} \tag{3.84}$$

If the Poisson model is correctly specified, $\mathrm{E}[(y_i - \lambda_i)^2/\lambda_i] = 1$, and thus $\mathrm{E}[\sum_{i=1}^{n}(y_i - \lambda_i)^2/\lambda_i] = n$. In practice, P is compared to $(n-k)$ in order to adjust for lost degrees of freedom due to estimation of λ_i. $P \neq n-k$ indicates a misspecification of the conditional mean or the distributional assumption.

An alternative goodness-of-fit statistic is the deviance

$$D = \sum_{i=1}^{n} \left\{ y_i \log\left(\frac{y_i}{\hat{\lambda}_i}\right) - (y_i - \hat{\lambda}_i) \right\} \tag{3.85}$$

(see McCullagh and Nelder, 1989). For the exponential Poisson model with intercept included, the sum over the second term on the right is zero, so that we can write

$$D = \sum_{i=1}^{n} y_i \log\left(\frac{y_i}{\hat{\lambda}_i}\right)$$

(where $y_i \log y_i = 0$ for $y_i = 0$). The deviance measures the difference between the maximum log-likelihood achievable and the log-likelihood achieved by the model under scrutiny. The deviance has an approximate chi-squared distribution with $n-k$ degrees of freedom.

Cameron and Windmeijer (1996) discuss the use of pseudo R-squared measures for determining goodness-of-fit within classes of count data regression models. They list desirable properties of pseudo R-squared measures:

1. $0 \leq R^2 \leq 1$.
2. R^2 does not decrease as regressors are added.
3. R^2 based on residual sum of squares coincides with R^2 based on explained sum of squares.
4. There is a correspondence between R^2 and a significance test on all slope parameters or on incremental slope parameters.

The preferred measure identified by Cameron and Windmeijer is one based on deviance residuals. For the Poisson regression model, it is given by

$$R^2 = \frac{\sum_{i=1}^{n} y_i \log(\hat{\lambda}_i/\bar{y})}{\sum_{i=1}^{n} y_i \log(y_i/\bar{y})} \tag{3.86}$$

Cameron and Windmeijer derive similar pseudo R-squared measures for the negative binomial regression model.

A final descriptive goodness-of-fit measure is used by Dionne and Vanasse (1992). They suggest to sum the individual predicted probabilities over all possible outcomes $0, 1, \ldots$, where for practical calculations some cutoff value has to be chosen (for instance, the maximum count observed in the sample, $m = \max_i(y_i)$). The summed probabilities for a specific outcome j

$$\hat{p}_j = \frac{1}{n} \sum_{i=1}^{n} \frac{e^{-\hat{\lambda}_i} \hat{\lambda}_i^j}{j!} \quad j = 0, 1, \ldots, m$$

can then be compared to the relative frequencies in the sample $\bar{p}_j$. Discrepancies indicate a poor performance of the model. A comparison of two models can be based on the proximity between average predicted probabilities and relative frequencies in the sample. Andrews (1988) shows how these fitted distributions can be subject to a formal chi-squared test.

3.5.4 Tests for Non-Nested Models

Standard statistical theory provides a wide range of tools for the testing of hypotheses which come in the form of parametric restrictions. A restriction transforms a general model G into a restricted model F. We say that F is nested in G. The restricted model can never be "better" than the general model, measured in terms of likelihood or coefficient of determination. Examples for two nested count data models are the Poisson and the negative binomial model, or the Poisson and the Poisson-log-normal model.

In count data modeling, one is often confronted with two models that are not nested in the above sense (for a general discussion of tests for non-nested models, see Pesaran, 1974). Two cases can be distinguished. In the first case, two models are *partially nested* or *overlapping* (Vuong, 1989). In this case neither of the two can be derived from the other through a parametric restriction on either model, and at the same time the models are identical for some joint restrictions.

This case is indeed very common for count data. For example, any two generalizations of the Poisson regression model (i.e., models that can be transformed to a Poisson model through a suitable parametric restriction) are pairwise overlapping. Since each of them can by definition be restricted to a Poisson model, they are equivalent if this restriction is imposed. Examples are the negative binomial model, the Poisson-log-normal model and the generalized Poisson distribution model. Similarly, hurdle Poisson and negative binomial

models overlap with zero inflated Poisson and negative binomial models. The equivalence of hurdle and zero inflated models in the case where there are constants only (i.e., all slope coefficients are restricted to zero) has been noted already by Mullahy (1986). A final example for two overlapping models is the Poisson model with log-linear mean function $\exp(x_i'\beta)$ and the Poisson model with linear mean function $x_i'\gamma$. Again, the two models are clearly distinct in the presence of genuine regressors. Yet, when slopes are zero and the constants obey the restriction $\exp(\beta_0) = \gamma_0$, the two models are the same.

Second, two models can be strictly non-nested. In this case, no set of restrictions is available that would render the two specifications formally equivalent. It is relatively hard to find an example for two strictly non-nested count data models. The only established example is the zero-inflated Poisson model with logit (or probit) parameterization of the probability of an excess zero, and the standard Poisson model. Here, the problem arises that the specific parameterization precludes that the extra probability takes the value 0 for any finite value of the model parameters. Hence, the two models cannot be equivalent.

There are two different ways of looking at models that are non-nested (strictly or overlapping). One is *hypothesis testing* and one is *model selection*. A hypothesis test addresses the issue whether the true conditional density $f_0(y|x)$ belongs to F or to G. By its very nature, it introduces an asymmetry between the null hypothesis and the alternative. To treat the models symmetrically, both models are considered consecutively under the null hypothesis. Combining the two tests, four outcomes are possible:

1. $H_0 = F$ is accepted and $H_0 = G$ is rejected.
2. $H_0 = F$ is rejected and $H_0 = G$ is accepted.
3. $H_0 = F$ is rejected and $H_0 = G$ is rejected.
4. $H_0 = F$ is accepted and $H_0 = G$ is accepted.

In situations (1) and (2), a coherent decision can be made. (1) leads to a decision in favor of F and (2) to a decision in favor of G. For (3) and (4), the results are conflicting. In (3) both models are rejected and in (4) the evidence cannot discriminate between the two models. For this reason an alternative approach, the *model selection* approach, considers situations where a decision in favor of one model has to be made. Also, model selection criteria are more suitable in situations where more than two models are considered. There, the hypothesis testing framework provides little guidance how to proceed.

The focus of this section, however, will be on hypothesis testing for two non-nested models. The existing results can be roughly divided into three approaches. The first generalizes the classical asymptotic tests, the second uses Monte Carlo simulations, and the third approach specifies and estimates a *hyper*-model.

Vuong Test

The extension of the likelihood ratio tests to situations of non nested models uses results on pseudo-true values minimizing the Kullback distance to the true conditional law to establish the asymptotic distribution of the test statistic. The corresponding results for the Wald and the Lagrange multiplier tests can be found in Gourieroux and Monfort (1989, Chap. 12). The likelihood ratio approach originates in Cox (1961) and has been extended by Vuong (1989). It is based on the observed difference

$$d_{\text{obs}} = \ell_f(\hat{\alpha}) - \ell_g(\hat{\beta}) \tag{3.87}$$

between the log-likelihood under model F and G evaluated at the maximum, respectively. Note that in contrast to nested models, this difference can be either negative or positive. The distribution of d_{obs} under any of the two models is, however, unknown. Cox (1961) derived a modified test statistic for a test of $H_0 = F_\alpha$ against $H_1 = G_\beta$:

$$T_f = \{\ell_f(\hat{\alpha}) - \ell_g(\hat{\beta})\} - \mathrm{E}_{\hat{\alpha}}[\ell_f(\hat{\alpha}) - \ell_g(\hat{\beta})] \tag{3.88}$$

where $\mathrm{E}_{\hat{\alpha}}[\cdot]$ is the expectation with respect to $F_{\hat{\alpha}}$. Due to the difficulty in calculating this expectation, the use of the Cox test is unrealistic in many practical situations.

Vuong (1989) has developed a considerably simpler test. His test statistic for non-nested models is

$$\mathrm{LR}_{\mathrm{NN}} = \frac{1}{\sqrt{n}}(\ell_f(\hat{\alpha}) - \ell_g(\hat{\beta}))/\omega \tag{3.89}$$

where

$$\omega^2 = \frac{1}{n}\sum_{i=1}^{n}(\ell_f(y_i|x_i,\hat{\alpha}) - \ell_g(y_i|x_i,\hat{\beta}))^2 - \left(\frac{1}{n}\sum_{i=1}^{n}\ell_f(y_i|x_i,\hat{\alpha}) - \ell_g(y_i|x_i,\hat{\beta})\right)^2$$

This test statistic is thus very simple to compute. All one needs to do is to divide the average difference of the log-likelihood functions (or the difference of the average), evaluated at the respective maximum likelihood estimates, by the estimated standard error of the average difference, which is just the standard deviation of the individual differences in log-likelihood divided by the square root of n.

None of the two models needs to be true. The test is aimed at selecting the model that is closest to the true conditional distribution. The null hypothesis is that the two models are equivalent:

$$H_0 : \mathrm{E}_0[\ell_f(\hat{\alpha}) - \ell_g(\hat{\beta})] = 0$$

Under the null hypothesis, the $\mathrm{LR}_{\mathrm{NN}}$ statistic converges in distribution to a standard normal distribution. To implement the test, let c denote the critical value for some significance level. If the statistic exceeds c, one rejects the null hypothesis (of equivalence) in favor of model f being better than model g. If the statistic is smaller than $-c$, one rejects the null hypothesis (of equivalence) in favor of model g being better than model f. Finally, if $|\mathrm{LR}_{\mathrm{NN}}| < c$ the null hypothesis is not rejected and one cannot discriminate between the two models given the data. In this case, both models are rejected. The possibility of rejecting both models is a conceptual weakness of this test, in which neither of the two models is true under H_0.

If the two models are overlapping – the standard situation in count data applications – the above test needs to be slightly modified. In particular, rather than applying it directly, a pre-test is required. In this pre-test, one needs to establish that the two models are not equivalent. In practice it is sufficient to separately perform t- or F-tests to see whether the parameter vectors violate the overlapping constraint, and the overlapping constraint can thus be rejected (see Vuong, 1989, particularly footnote 6).

For example, when testing the negative binomial model against the Poisson-log-normal model, this would amount to testing whether the two dispersion parameters are individually significantly greater than zero. When testing the hurdle Poisson model against the zero inflated Poisson model, it is sufficient that the slope vectors in both models are significantly different from zero, so that one rejects the null hypothesis of two constant index functions, the requirement for the overlap (Mullahy, 1986).

It should be noted that, in practice, many researchers ignore this issue, assuming right away that the overlapping models are not equivalent, and conducting the (nonnested) likelihood ratio test directly without pretesting.

Simulation-Based Tests

A second approach for testing non-nested hypotheses is simulation-based. This test has been put forward by Williams (1970). Essentially, a large number of data sets is generated under each of the two models. The models are re-estimated and a likelihood ratio is calculated. Denote by d_f (d_g) the distribution of the (log of the) likelihood ratio under F (G). Comparing the observed likelihood ratio with d_f (d_g) then provides evidence in favor of F_α or in favor of G_β. The following four steps can be distinguished:

- Obtain estimates $\hat{\alpha}$ and $\hat{\beta}$ and calculate the observed log-likelihood difference d_{obs}
- Simulate R sets of endogenous variables y_r under F_α with $\alpha = \hat{\alpha}$ and identical x. Then re-estimate each conditional model with y_r $i = 1, \ldots R$ and x to obtain $\hat{\alpha}_{fr}$ and $\hat{\beta}_{fr}$ and calculate $d_{fr} = \ell_f(\hat{\alpha}_{fr}) - \ell_g(\hat{\beta}_{fr})$.
- Simulate R sets of endogenous variables y_r under G_β with $\beta = \hat{\beta}$ and identical x. Then re-estimate each conditional model with y_r $i = 1, \ldots R$ and x to obtain $\hat{\alpha}_{gr}$ and $\hat{\beta}_{gr}$ and calculate $d_{gr} = \ell_f(\hat{\alpha}_{gr}) - \ell_g(\hat{\beta}_{gr})$.

- Compare the value d_{obs} with the empirical distribution of d_{fr} and of d_{gr} to provide evidence whether the observed log-likelihood difference is more compatible with model F or with model G.

As a result, none of the simulations may generate values close to the observed ones, the simulations may support a particular model, or they may not be able to discriminate between the two models.

The question arises, how the y_r can be simulated. Williams (1970) proposed *parametric simulation*. In case of the Poisson models with different mean functions, call them λ_f and λ_g, this amounts to repeated draws from Poisson distributions with means fixed at $\bar{\lambda}_f = \lambda_f(\hat{\alpha})$ $\bar{\lambda}_g = \lambda_g(\hat{\beta})$, respectively. Alternatively, Hinde (1992) suggests a non-parametric bootstrap simulation, i.e., re-sampling from the observed residuals.

Artificial Nesting

A third method for testing non-nested models is the construction of a hypermodel (See for instance Gourieroux and Monfort, 1989, Chap. 12). In general, hyper-models contain an additional parameter, and a test between two models comes down to a test of a restriction on the hyper-parameter. One important example in the count data literature, the Negbin$_k$ model, is presented in Chap. 4.3. It can be used to discriminate between the two common specifications of the negative binomial model, Negbin I and Negbin II. The general remarks about the coherence of the test for non-nested hypotheses apply here as well: The test may produce a conclusive result, or the evidence may be inconclusive, either rejecting both or none of the models.

Hausman Test

In certain situations, two non-nested models can be tested by way of a Hausman test (Hausman, 1978). The underlying test idea is quite general and not restricted to maximum likelihood estimation, nor to count data models. Assume that under the null hypothesis, model 1 gives an estimator that is both consistent and efficient. The alternative model 2 gives an estimator that is consistent but inefficient. Further, assume that under the alternative, model 1 is inconsistent but model 2 remains consistent. The Hausman test is based on the distance

$$HT = (\hat{\theta}_1 - \hat{\theta}_2)'[\text{Var}(\hat{\theta}_1 - \hat{\theta}_2)]^{-1}(\hat{\theta}_1 - \hat{\theta}_2) \tag{3.90}$$

Under H_0, both estimators are consistent. Hence, $\hat{\theta}_1$ and $\hat{\theta}_2$ should be similar. Under the alternative, $\hat{\theta}_1$ is inconsistent. Thus, "large" values of $\hat{\theta}_1 - \hat{\theta}_2$ lead to a rejection of H_0. The computation of the test statistic requires estimation of $\text{Var}(\hat{\theta}_1 - \hat{\theta}_2)$. Since $\hat{\theta}_1$ and $\hat{\theta}_2$ use the same data they could be correlated which would cause considerable difficulties.

However, recall that model 1 is by assumption efficient under the null hypothesis. It can be shown (Hausman, 1978) that the asymptotic covariance of an efficient estimator with its difference from an inefficient estimator must be zero:

$$\mathrm{Cov}(\hat{\theta}_1, \hat{\theta}_1 - \hat{\theta}_2) = \mathrm{Var}(\hat{\theta}_1) - \mathrm{Cov}(\hat{\theta}_1, \hat{\theta}_2) = 0$$

and hence $\mathrm{Cov}(\hat{\theta}_1, \hat{\theta}_2) = \mathrm{Var}(\hat{\theta}_1)$. Thus, under H_0,

$$\begin{aligned} \mathrm{Var}(\hat{\theta}_1 - \hat{\theta}_2) &= \mathrm{Var}(\hat{\theta}_1) - 2\mathrm{Cov}(\hat{\theta}_1, \hat{\theta}_2) + \mathrm{Var}(\hat{\theta}_2) \\ &= \mathrm{Var}(\hat{\theta}_2) - \mathrm{Var}(\hat{\theta}_1) \end{aligned} \tag{3.91}$$

The resulting test statistic has an asymptotic chi-squared distribution with q degrees of freedom, where q is the dimension of $\hat{\theta}_1$.

In count data modeling, an application of this test occurs with panel data, where the random effects Poisson model is tested against the fixed effects Poisson model. Implicitly, one tests the assumption that the individual specific unobserved error component is uncorrelated with the regressors. A related application arises in tests for endogeneity of regressors.

3.6 Outlook

This chapter discussed various sources of, and tests for, misspecification. The next chapters introduce alternative count data models that overcome the restrictiveness of the Poisson specification. While this chapter has explored a substantial number of potential deficiencies, a useful classification of the received literature can be based on a tripart classification:

- Unobserved heterogeneity (Chapter 4)
- Selectivity and Endogeneity (Chapter 5)
- Extra Zeros (Chapter 6)

The problem of unobserved heterogeneity occupies a somewhat special place, as it is the only problem among the three that does not necessarily require methods beyond the standard Poisson model. The pseudo likelihood method was discussed before. Thus, the issue in Chapter 4 is rather one of efficient estimation, making most of the information available in the data. Such efficiency gains can be obtained by modeling the unobserved heterogeneity, either by making parametric assumption on its distribution, or alternatively by considering less restrictive semi-parametric assumptions, or even non-parametric estimation of the unknown distribution.

Unobserved heterogeneity can give rise to endogeneity, if the unobservables are not independent of the regressors of interest. This is discussed in Chapter 5, together with a number of models for non-random sampling designs.

The general thrust of the arguments in Chapters 4 and 5 are very much reminiscent of standard methods known for the linear model. For example,

unobserved heterogeneity, and the resulting problem of overdispersion, gives raise to a trade-off between efficiency and robustness that is essentially identical to the trade-off caused by heteroskedasticity in the linear model, when contemplating the modeling option of generalized least squares versus ordinary least squares with White-Huber standard errors. The methods for selectivity and endogeneity are also largely borrowed from the literature on linear models and adjusted to the exponential regression model.

In contrast to that, the chapter on extra zeros has no direct equivalence in the linear model literature. It is different because it takes the nature of the dependent variable seriously – positive probability mass at discrete outcomes – and studies various flexible probability models, and conditions under which these models allow for meaningful inferences on distributional shifts, i.e., marginal probability effects.

Whenever possible and appropriate, models are represented not as simple "data fitting tools", but rather as manifestations of an underlying structural data generating process (such as a stochastic count process), which in turn means that the parameters offer interpretations of these underlying parameters. Finally, it should be noted that these three chapters maintain the assumption of cross-section data – models for correlated count data are discussed later on in Chapter 7.

4

Unobserved Heterogeneity

4.1 Introduction

4.1.1 Conditional Mean Function

This chapter discusses mixture models for unobserved heterogeneity. The problem of unobserved heterogeneity arises if the explanatory variables do not account for the full amount of individual heterogeneity in the conditional mean of the dependent variable. Assume that the true model is

$$\mathrm{E}(y|x,z) = \exp(x'\beta + w'\gamma)$$

However, w is unobserved by the econometrician. Thus, we have instead

$$\mathrm{E}(y|x,v) = \exp(x'\beta + v) \tag{4.1}$$

where $v = w'\gamma$ represents *unobserved heterogeneity*. An alternative, equivalent representation of the model uses $u = \exp(v)$, so that

$$\mathrm{E}(y|x,u) = \exp(x'\beta)u$$

Can there be any hope that the Poisson maximum likelihood estimator based on the mean function $\lambda = \exp(x'\beta)$ maintains some or all of its usual desirable properties?

It is useful to recall the results for omitted variables in the linear regression model. There, bias arises whenever x and w are correlated. However, without correlation, omitted variables do not cause a problem. In fact, omitted variables are, apart from measurement error in y, a standard argument for introducing a stochastic relation with additive error term to begin with.

This basic line of reasoning carries over to the problem of omitted variables in the Poisson regression. The main difference is that, because of the nonlinearity of the mean function, zero correlation is not sufficient and we have to make the stronger assumption of mean independence. In particular, we will assume that

$$\mathrm{E}[\exp(v)|x] = \mathrm{E}[\exp(v)] \tag{4.2}$$

is a constant and not a function of x. Note that it is neither sufficient nor necessary to assume that $\mathrm{E}(v|x) = \mathrm{E}(v)$. For instance, if $\mathrm{E}(v|x) = \mathrm{E}(v)$ and the variance of v is a function of x, it follows that $\mathrm{E}[\exp(v)|x]$ is a function of x due to the exponential transformation. We could assume full independence between x and v, but that assumption is unnecessarily strong. What we need is exactly (4.2). In the following, it will prove useful to keep track of the constant term in the linear predictor $x'\beta$. Let the constant be α so that we can write the base model with unobserved heterogeneity as

$$\mathrm{E}(y|x, v) = \exp(\alpha + x'\beta + v) \tag{4.3}$$

which yields, using (4.2), the mean function conditional on x, but unconditional on v,

$$\mathrm{E}(y|x) = \exp(\alpha + x'\beta)\mathrm{E}[\exp(v)|x] = \exp(\tilde{\alpha} + x'\beta) \tag{4.4}$$

where $\tilde{\alpha} = \alpha + \log \mathrm{E}[\exp(v)]$. Of course, if we *assume* that $\mathrm{E}[\exp(v)] = 1$, a common normalization, then $\alpha = \tilde{\alpha}$. In this case, we see that unobserved heterogeneity does not change the conditional expectation function at all.

4.1.2 Partial Effects with Unobserved Heterogeneity

What are the partial effects of interest in a model with unobserved heterogeneity? From (4.3), we can define

$$\frac{\partial \mathrm{E}(y|x, v)}{\partial x_j} = \exp(\alpha + x'\beta + v)\beta_j \tag{4.5}$$

This is the *ceteris paribus* effect of x on the expected y, keeping constant the heterogeneity term v. Unlike in the linear model with additive effects, the multiplicative model implies that partial effects are a function of the unobserved heterogeneity. One approach would be to evaluate these partial effects at a representative value, say $v = 0$ (although this value may be representative only for a small fraction of the population as, for continuous v, $P(v = 0) = 0$). Thus

$$\frac{\partial \mathrm{E}(y|x, v = 0)}{\partial x_j} = \exp(\alpha + x'\beta)\beta_j$$

The problem with this approach is, however, that the partial effect is not identified. Recall from (4.4) that the constant of the mean function with unobserved heterogeneity is $\tilde{\alpha}$, not α as required. Unless one makes the arbitrary identifying assumption that $\mathrm{E}(u) = 1$, i.e. $\alpha = \tilde{\alpha}$, the conditional partial effects are undetermined.

There are two ways out of this conundrum. The first one, advocated by Wooldridge (2002), is to focus on *average partial effects*. Taking expectations of (4.5) with respect to v, we obtain

$$\mathrm{E}_v\left(\frac{\partial \mathrm{E}(y|x, v)}{\partial x_j}\right) = \exp(\alpha + x'\beta)\mathrm{E}(\exp(v))\beta_j = \exp(\tilde{\alpha} + x'\beta)\beta_j$$

which is identified from the mean function $\mathrm{E}(y|x)$. Taking expectations works well because of the multiplicative separability of the unobserved heterogeneity component.

The second approach, and possibly the more natural one, given the multiplicative model, is to focus on *relative partial effects* instead. Clearly,

$$\frac{\partial \mathrm{E}(y|x,v)/\mathrm{E}(y|x,v)}{\partial x_j} = \beta_j$$

and β_j measures the proportional effect of x_j both the conditional expectation $\mathrm{E}(y|x,v)$ and the unconditional expectation $\mathrm{E}(y|x)$.

4.1.3 Unobserved Heterogeneity in the Poisson Model

Let $y|x,u$ have a Poisson distribution with conditional mean and variance

$$\mathrm{E}(y|x,u) = \mathrm{Var}(y|x,u) = \exp(x'\beta)u,$$

where u is distributed mean-independently of x with $\mathrm{E}(u|x)$ normalized to 1. Also assume that the variance of u, σ_u^2, is a constant not depending on x. This assumption is not really important. It only serves to obtain benchmark results that are easily modified for the case where σ_u^2 is a function of x.

The distribution of y, marginalized with respect to u but conditional on x has then expectation

$$\mathrm{E}(y|x) = \exp(x'\beta)\mathrm{E}(u|x) = \exp(x'\beta)$$

and variance

$$\mathrm{Var}(y|x) = \mathrm{E}_u[\mathrm{Var}(y|x,u)] + \mathrm{Var}_u[\mathrm{E}(y|x,u)] = \exp(x'\beta) + \sigma_u^2[\exp(x'\beta)]^2$$

an application of the variance decomposition theorem, where we use that $\mathrm{E}(y|x,u) = \mathrm{Var}(y|x,u)$ and $\mathrm{E}(u) = 1$. Therefore, $\mathrm{Var}(y|x) > \mathrm{E}(y|x)$, i.e., unobserved heterogeneity of this form causes overdispersion of the conditional model for $y|x$ relative to the Poisson model. This is an important result.

Thus, if one estimates a Poisson model in the presence of unobserved heterogeneity of the type discussed here, the model is misspecified. Importantly, however, the mean function is correctly specified, which means that one can apply the results for Pseudo-Maximum Likelihood estimation from Chapter 3.3 and obtain consistent parameter estimates and valid inference based on the Poisson model. This a result of great practical relevance. It suggests that Poisson regression, using robust standard errors, is entirely appropriate despite of unobserved heterogeneity. As unobserved heterogeneity can almost never be ruled out – testing almost always rejects the null hypothesis of no overdispersion – this is a very useful feature of the Poisson model.

4.1.4 Parametric and Semi-Parametric Models

Of course, one might instead be interested in efficient estimation in a model with unobserved heterogeneity. In this case, we require the marginal distribution $f(y|x)$ which is obtained by taking the expectation of the conditional distribution $f(y|x,u)$ (which is in our case a Poisson distribution) with respect to u:

$$f(y|x) = \int_0^\infty f(y|x,u)g(u|x)\,du \tag{4.6}$$

For instance, if $f(y|x,u)$ is of the Poisson form, we obtain

$$f(y|x) = \frac{\lambda^y}{y!}\int e^{-\lambda u}u^y g(u|x)\,du \tag{4.7}$$

This approach requires additional assumptions, as we need to specify $g(u|x)$, the distribution of the unobserved heterogeneity. Going with the assumption of independence, we have that $g(u|x) = g(u)$. Moreover, we know that u has to be non-negative, a restriction that should be reflected in the selection of $g(u)$. Candidate distributions for u in the literature are the gamma, the log-normal, and the inverse Gaussian distributions. The resulting fully parametric mixture models are discussed in greater detail in Chapter 4.

The gain of introducing a parametric assumption can be an increase in efficiency. The downside is a potential loss of consistency if the specific parametric assumption is wrong. More robust methods are obtained by semiparametric methods that approximate the distribution of u, either using Laguerre polynomials and moment generating functions (Gurmu, Rilstone, and Stern, 1998), or discrete factor approximations (Brännäs and Rosenqvist, 1994). In either case, some efficiency is lost but substantial robustness may be gained from weaker assumptions. Both approaches are discussed later on.

A third, from an estimation standpoint entirely different, situation is encountered if repeated measurements are available for the same individual, that is, if panel data are available. Methods to control for unobserved heterogeneity in panel count data models are presented in Chap. 7.2.

4.2 Parametric Mixture Models

Different Poisson mixture models can be distinguished depending on the specific assumptions made on the distribution of u. Johnson and Kotz (1969, Chap. 8) discuss a variety of mixing distributions. Another general reference is Karlis and Xekalaki (2005). The choice is determined by the requirement that u be non-negative, and applications in regression analysis so far have concentrated on three distributions: the gamma distribution, the inverse Gaussian distribution, and the log-normal distribution.

4.2.1 Gamma Mixture

The earliest mention of the Poisson-gamma mixture appears to be by Greenwood and Yule (1920). A random variable u is gamma distributed $\Gamma(\alpha, \beta)$ if the density takes the form

$$g(u; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} u^{\alpha-1} e^{-\beta u} \tag{4.8}$$

Mean and variance are $\mathrm{E}(u) = \alpha/\beta$ and $\mathrm{Var}(u) = \alpha/\beta^2$, respectively. Let $\alpha = \beta$, a restriction that reduces the number of free parameters from two to one. Then $\mathrm{E}(u) = 1$ and $\mathrm{Var}(u) = \alpha^{-1}$. Recall that $\tilde{\lambda} = \lambda u$. Applying the change of variable technique, $\tilde{\lambda}$ has a gamma distribution

$$\begin{aligned} h(\tilde{\lambda}; \lambda, \alpha) &= \frac{\alpha^\alpha}{\Gamma(\alpha)} \left(\frac{\tilde{\lambda}}{\lambda}\right)^{\alpha-1} e^{-\frac{\tilde{\lambda}\alpha}{\lambda}} \frac{1}{\lambda} \\ &= \frac{(\alpha/\lambda)^\alpha}{\Gamma(\alpha)} \tilde{\lambda}^{\alpha-1} e^{-\tilde{\lambda}\frac{\alpha}{\lambda}} \end{aligned} \tag{4.9}$$

with mean λ and variance $\alpha^{-1}\lambda^2$. The gamma distribution is a *scale* family, i.e., it is closed under scale transformations.

As demonstrated in Chap. 2.5.1, integration (3.60) of a Poisson-gamma mixture leads to the negative binomial distribution for y:

$$f(y|\alpha, \lambda) = \frac{\Gamma(\alpha + y)}{\Gamma(\alpha)\Gamma(y+1)} \left(\frac{\alpha}{\lambda + \alpha}\right)^\alpha \left(\frac{\lambda}{\lambda + \alpha}\right)^y \tag{4.10}$$

with

$$\mathrm{E}(y|\alpha, \lambda) = \lambda$$

and

$$\mathrm{Var}(y|\alpha, \lambda) = \lambda + \alpha^{-1}\lambda^2 .$$

The regression model is completed by setting $\lambda_i = \exp(x_i'\beta)$. An alternative common specification uses $\alpha = \sigma^{-2}\lambda_i$, resulting in a negative binomial model with linear variance function. Because of its importance and widespread use in applied work, this model and its various parameterizations are discussed in greater detail below.

4.2.2 Inverse Gaussian Mixture

The Poisson-inverse Gaussian mixture model was discussed, among others, by Dean, Lawless, and Willmot (1989). Early applications tended to focus on the univariate case. A regression parameterization that is directly comparable to the negative binomial regression model was used by Guo and Trivedi (2002)

who included this model in a comparative performance study in an application to patent data. As before, the marginal probabilities are obtained after integration

$$f(y) = \int_0^\infty f(y|\tilde{\lambda})g(\tilde{\lambda})\, d\tilde{\lambda} \tag{4.11}$$

Let $\tilde{\lambda}$ be *inverse Gaussian* distributed with density

$$g(\tilde{\lambda}) = \sqrt{\frac{\alpha}{2\pi\tilde{\lambda}^3}} \exp\left(-\frac{\alpha(\tilde{\lambda}-\lambda)^2}{2\tilde{\lambda}\lambda^2}\right) \tag{4.12}$$

where $\alpha > 0$, $\tilde{\lambda} > 0$ and $\lambda > 0$. The inverse Gaussian distribution has mean λ and variance λ^3/α. If we parameterize $\lambda_i = \exp(x_i'\beta)$, the conditional mean and conditional variance of the Poisson-inverse Gaussian distribution are

$$\mathrm{E}(y_i|x_i) = \exp(x_i'\beta)$$

and

$$\mathrm{E}(y_i|x_i) = \mathrm{E}(y_i|x_i) + \alpha^{-1}[\mathrm{E}(y_i|x_i)]^3$$

(see Guo and Trivedi, 2002).

Though the integration (4.11) does not yield a closed form for this choice of g, the probability generating function can be calculated using the methods introduced in Appendix A. Dean, Lawless, and Willmot (1989) note that the probabilities of the mixture distribution can be calculated recursively using a second order difference equation. They also derive analytical first and second derivatives of the count data log-likelihood.

4.2.3 Log-Normal Mixture

In the log-normal mixture, we assume that

$$\lambda = \exp(x'\beta + v)$$

where v has a normal distribution with mean $-\sigma^2/2$ and variance σ^2. The normal distribution appears to be an immensely sensible choice, because if there are many independent unobserved factors, then the sum of them may converge to a normal distribution by virtue of a central limit theorem.

If v is normal distributed as above, then $u = \exp(v)$ is log-normal distributed with mean equal to one and variance equal to $\mathrm{Var}(u_i) = e^{\sigma^2} - 1$. The probability function of the Poisson-log-normal model can be written as

$$f(y|x,v) = \frac{\exp(-\exp(x'\beta + v))\exp(x'\beta + v)^y}{y!}$$

where $v \sim N(0, \sigma^2)$, i.e.,

$$f(v) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{v}{\sigma}\right)^2}$$

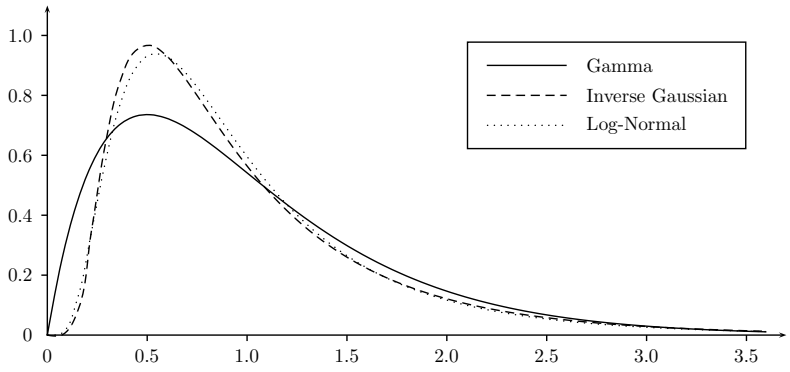
Therefore, the observed data distribution function is obtained after integration as

$$f(y|x) = \int_{-\infty}^{\infty} \frac{\exp(-\exp(x'\beta + v))\exp(x'\beta + v)^y}{y!} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{v}{\sigma}\right)^2} dv \quad (4.13)$$

No closed form solution of this integral is available for the Poisson-log-normal model. Hinde (1982) shows how maximum likelihood estimates can be obtained using a combination of numerical integration, the EM algorithm and iteratively reweighed least squares. With modern computing power, direct computation by Gauss-Hermite quadrature is quite straightforward, and maximum likelihood estimation in most situations is as fast as estimation of the negative binomial model. Details of the Gauss-Hermite procedure for this model are provided in Appendix B.

Relative to the Poisson model, the three models presented here share the common feature that they introduce one additional parameter each, essentially a variance parameter. This parameter can be specified such as to preserve identical mean and variances for the distribution of u, and thus, for the mixture models for y. The three models have the same variance function as long as $e^{\sigma^2} - 1 = \alpha^{-1}$, or, equivalently, $\sigma_u^2 = \log(1 + \alpha) - \log(\alpha)$, where α^{-1} is the variance of the multiplicative heterogeneity factor in the inverse Gaussian and gamma mixture models. The models are thus distinguished by their higher moments. Fig. 4.1 gives three density functions with $\mathrm{E}(u) = 1$ and $\mathrm{Var}(u) = 0.5$. We see that the density functions of the log-normal and inverse Gaussian models are virtually identical, whereas the density function of the Gamma distribution differs.

Fig. 4.1. Probability Density Functions of Gamma, Inverse Gaussian, and Log-Normal Distributions



One of the practical advantages of the Poisson-log-normal model is that it is readily extended to the multivariate case (Aitchison and Ho, 1989; Chib and Winkelmann, 2001). Moreover, it has a natural interpretation. Assume that, as in the linear model, the error $v = \log u$ captures the effect of several additive omitted variables. As mentioned before, if there are many omitted factors, and if these factors are independent, then central limit theorems can be invoked in order to establish normality of v. This model is not only appealing from a theoretical perspective; results in the application section of this book show that it fits the data often much better than the negative binomial model. These results suggest that the previous neglect of the Poisson-log-normal model in the literature should be reconsidered in future applied work.

4.3 Negative Binomial Models

The negative binomial distribution is the most commonly used alternative to the Poisson model when it is doubtful whether the strict requirements of independence of the underlying process, and inclusion of all relevant regressors, are satisfied. In particular, the negative binomial (Negbin) model is appropriate when the conditional distribution of $y|\tilde{\lambda}$ is Poisson distributed and $\tilde{\lambda}$ is independently gamma distributed. Thus, the Negbin model has the interpretation of a Poisson mixture model that accounts in a specific way for the randomness of the Poisson parameter $\tilde{\lambda}$. Alternatively, the Negbin model arises when the underlying count process is not independent, and when the dependence can be described through a specific type of *true contagion* (See Chap. 2.2.5). Further references on the Negbin model include Cameron and Trivedi (1986), Lawless (1987b) and Hausman, Hall and Griliches (1984).

The probability function of the negative binomial model has been given in (4.10). To make the step to the Negbin regression model, the parameters α and λ are specified in terms of exogenous variables.

The Negbin II model is obtained for $\alpha = \sigma^{-2}$ and $\lambda = \exp(x'\beta)$. In this case, the conditional expectation function is

$$\mathrm{E}(y|x) = \exp(x'\beta) \tag{4.14}$$

while the conditional variance function is given by

$$\mathrm{Var}(y|x) = \exp(x'\beta) + \sigma^2[\exp(x'\beta)]^2 \tag{4.15}$$

The conditional variance is always greater than the conditional mean: the negative binomial model is a model for *overdispersion*. The Negbin I model is obtained by letting α vary across individuals such that $\alpha = \sigma^{-2}\exp(x'\beta)$ and $\lambda = \exp(x'\beta)$. This parameterization produces a variance that is a linear function of the mean:

$$\mathrm{Var}(y|x) = (1 + \sigma^2)\exp(x'\beta) \tag{4.16}$$

Another way of characterizing the difference between the Negbin I and Negbin II models is in terms of a dispersion function ϕ, such that $\mathrm{Var}(y|x) =$

$\phi \mathrm{E}(y|x)$. For the Negbin I model, $\phi = (1 + \sigma^2)$, a constant function, whereas for the Negbin II model, $\phi = 1 + \sigma^2 \exp(x'\beta)$.

4.3.1 Negbin II Model

The (conditional) probability function of the Negbin II model can be written as

$$f(y|\cdot) = \frac{\Gamma(\sigma^{-2} + y)}{\Gamma(\sigma^{-2})\Gamma(y+1)} \left(\frac{\sigma^{-2}}{\exp(x'\beta) + \sigma^{-2}}\right)^{\sigma^{-2}} \left(\frac{\exp(x'\beta)}{\exp(x'\beta) + \sigma^{-2}}\right)^{y}$$

For $\sigma^2 \to 0$, this model converges to the Poisson regression model (See Chap. 2.3.1). Since $\sigma^2 \geq 0$ the Poisson model is obtained at the boundary of the parameter space. This has to be kept in mind when evaluating the model: a modified likelihood ratio test has to be used to test $H_0 : f$ *is Poisson* against $H_1 : f$ *is negative binomial*. The problem of testing for restrictions at the boundary of the parameter space was discussed in Chap. 3.5.1.

Assuming an independent sample, the log-likelihood function of the Negbin II model is given by

$$\begin{aligned}\ell(\beta, \sigma^2) = \sum_{i=1}^{n} \Bigg[&\left(\sum_{j=1}^{y_i} \log(\sigma^{-2} + j - 1) \right) - \log y_i! \\ &-(y_i + \sigma^{-2}) \log(1 + \sigma^2 \exp(x_i'\beta)) + y_i \log \sigma^2 + y_i x_i'\beta \Big] \end{aligned} \tag{4.17}$$

where the ratio of gamma functions in the first line was simplified with the help of equation (2.45). The Negbin II maximum likelihood estimators $\hat{\beta}$ and $\hat{\sigma}^2$ are obtained as solutions to the first-order conditions

$$\sum_{i=1}^{n} \frac{y_i - \exp(x_i'\beta)}{1 + \sigma^2 \exp(x_i'\beta)} x_i = 0 \tag{4.18}$$

and

$$\begin{aligned}\sum_{i=1}^{n} \Bigg[&\frac{1}{\sigma^4} \left(\log(1 + \sigma^2 \exp(x_i'\beta)) - \sum_{j=1}^{y_i} \frac{1}{\sigma^{-2} + j - 1} \right) \\ &- \frac{(y_i + \sigma^{-2}) \exp(x_i'\beta)}{1 + \sigma^{-2} \exp(x_i'\beta)} + \frac{y_i}{\sigma^2} \Bigg] = 0 \end{aligned} \tag{4.19}$$

Moreover, it can be shown (See Lawless, 1987b) that the information matrix is block-diagonal. Therefore, $\hat{\sigma}^2$ and $\hat{\beta}$ are asymptotically independent. The variance of $\hat{\beta}$ is given by

$$\mathrm{Var}(\hat{\beta}) = \left(\sum_{i=1}^{n} \frac{\exp(x_i'\beta)}{1 + \sigma^2 \exp(x_i'\beta)} x_i x_i' \right)^{-1} \tag{4.20}$$

4.3.2 Negbin I Model

The Negbin I model has log-likelihood function

$$\ell(\beta,\sigma^2)=\sum_{i=1}^{n}\left[\left(\sum_{j=1}^{y_i}\log(\sigma^{-2}\exp(x_i'\beta)+j-1)\right)-\log y_i!\right.$$
$$\left.-(y_i+\sigma^{-2}\exp(x_i'\beta))\log(1+\sigma^2)+y_i\log\sigma^2\right] \tag{4.21}$$

with first-order conditions for $\hat{\beta}$:

$$\sum_{i=1}^{n}\left[\left(\sum_{j=1}^{y_i}\frac{\sigma^{-2}\exp(x_i'\beta)}{\sigma^{-2}\exp(x_i'\beta)+j-1}\right)x_i+\sigma^{-2}\exp(x_i'\beta)x_i\right]=0 \tag{4.22}$$

In contrast to the Negbin II model, the first-order conditions of the Negin I model are not of the form $\sum(y_i-\mu_i)f(\mu_i)=0$. The Negbin I model does not fall within the class of linear exponential families, and the robustness results derived in Chap. 3.3 therefore do not apply in this case. In fact, the Negbin II model is the only model in that family. Relatedly, it is also the only Negbin model with block-diagonal information matrix.

4.3.3 Negbin$_k$ Model

Despite these advantages of the Negbin II model, one might nevertheless wish to embark on a search for alternative estimators that are asymptotically efficient if correctly specified. One such model is the generalized negative binomial model of Winkelmann and Zimmermann (1991, 1995). A similar model has been employed independently by Ruser (1991). This model was re-discovered by Saha and Dong (1997) who apparently were unaware of the previous literature. Let $\alpha=\sigma^{-2}\lambda^{1-k}$ and $\lambda=\exp(x'\beta)$. k is a continuous *non-linearity* parameter. Compared to the Poisson model, two additional parameters have to be estimated and this model has been called Negbin$_k$.

The Negbin$_k$ can be interpreted as a hyper-model for the non-nested Negbin I and Negbin II models. In particular, the Negbin$_k$ nests the Negbin II and Negbin I through the parametric restrictions $k=1$ and $k=0$, respectively. Thus, a test between the two non-nested sub-models can proceed as described in Chap. 3.5.4. (See Ozuna and Gomez (1995) for a number of other approaches for testing between the Negbin I and Negbin II models.)

One possible representation of the probability function of the Negbin$_k$ model makes use of the following notation. First,

$$\left(\frac{\alpha}{\lambda+\alpha}\right)^{\alpha}=(\alpha^{-1}\lambda+1)^{-\alpha}$$
$$=(\sigma^2\lambda^k+1)^{-\lambda^{1-k}/\sigma^2}$$

Moreover,

$$\left(\frac{\lambda}{\lambda+\alpha}\right)^{y} = (1+\alpha\lambda^{-1})^{-y}$$
$$= \prod_{i=1}^{y} \frac{1}{1+\sigma^{-2}\lambda^{-k}}$$

Finally, using the recursive property of the gamma function,

$$\frac{\Gamma(\alpha+y)}{\Gamma(\alpha)\Gamma(y)} = \prod_{j=1}^{y} \frac{\sigma^{-2}\lambda^{1-k}+j-1}{j}$$

Putting everything together, the probability function of the Negbin_k can be expressed as

$$f(y|\lambda,\sigma^2,k) = C \times \begin{cases} \prod_{j=1}^{y} \frac{\sigma^{-2}\lambda^{1-k}+j-1}{(1+\sigma^{-2}\lambda^{-k})j} & \text{for } y = 1,2,\ldots \\ 1 & \text{for } y = 0 \end{cases} \tag{4.23}$$

with

$$C = (\sigma^2\lambda^k+1)^{-\lambda^{1-k}/\sigma^2}$$
$$\lambda = \exp(x'\beta)\,, \sigma^2 \geq 0.$$

Given an independent sample of observations, the log-likelihood of the sample is the sum of the logarithm of the probabilities $f(y)$.

4.3.4 Negbin$_X$ Model

Yet another parameterization of the negative binomial distribution is proposed by Santos Silva and Windmeijer (2001). Remember from Chap. 2.5.2 that the negative binomial model can be represented as a stopped sum (or compound) distribution, where

$$Y = R_1 + R_2 + \ldots + R_S = \sum_{i=j}^{S} R_j$$

where $S = 0,1,2,\ldots$ is Poisson distributed, and the components $R_j = 1,2,\ldots$ are identically and independently logarithmic distributed. The logarithmic distribution has a single parameter θ with $0 < \theta < 1$. So it is natural to allow for covariates by letting

$$\theta = \frac{\exp(x'\gamma)}{1+\exp(x'\gamma)}$$

Accordingly,

$$\frac{\theta}{1-\theta} = \exp(x'\gamma)$$

From the properties of the logarithmic distribution, it follows that the expected number of counts in each component is

$$\mathrm{E}(R_j) = \frac{\exp(x'\gamma)}{\log[1+\exp(x'\gamma)]}$$

If one lets for the Poisson part in addition

$$\lambda = \mathrm{E}(S) = \exp(x'\beta)$$

as usual, then it follows that Y is negative binomial distributed with parameters $\alpha = \exp(x'\beta)/\log[1+\exp(x'\gamma)]$ and $\lambda = \exp(x'\beta)$. Substituting these expressions into the negative binomial probability function (2.36), and after some further simplifications, one obtains the Negbin_X probability function

$$f(y|x) = \frac{\Gamma\left(y + \frac{\exp(x'\beta)}{\log[1+\exp(x'\gamma)]}\right)\exp(-\exp(x'\beta))}{\Gamma(y+1)\Gamma\left(\frac{\exp(x'\beta)}{\log[1+\exp(x'\gamma)]}\right)(1+\exp(-x'\gamma))^y} \tag{4.24}$$

with

$$\mathrm{E}(y|x) = \frac{\exp(x'\beta + x'\gamma)}{\log[1+\exp(x'\gamma)]}\,.$$

Of course, one can modify the model further by including different sets of regressors z and x in the different parts of the model. Usually, there will be little a-priori reason to justify such a selection, however, and the model will include two coefficients for each available covariate. The interesting aspect of the model is the interpretation of the underlying data generating process. The overall effect of a regressor on the total number of counts is the sum of two separate effects. First, a variable may affect the number of components S. Second, a variable may affect the number of counts in each component R_j. This separation may have analogies in real life processes. Santos Silva and Windmeier motivate their model by the demand for doctor visits. Here, the total number of visits may depend on the total number of sickness spells over a period plus the number of visits within each spell.

Estimation

A remarkable result, due to Holgate and restated in Guo and Trivedi (2002) is that all continuous mixtures based on the Poisson distibution – this includes all three models discussed here, Negbin, Poisson inverse gamma and Poisson log-normal – have unimodal likelihood functions. Hence, applications of standard Newton-Raphson or BFGS algorithms will find the global maximum of the log-likelihood function.

4.4 Semiparametric Mixture Models

4.4.1 Series Expansions

Gurmu, Rilstone, and Stern (1998) develop a semiparametric estimation approach for overdispersed count regression models based on series expansions

for the unknown density of the unobserved heterogeneity component. They notice that while conventional approaches to unobserved heterogeneity impose *ad-hoc* restrictions on the functional form of the mixing distribution whose violation causes the estimator to be inconsistent, quasi-likelihood methods do not use information on higher order moments and hence are inefficient. Furthermore, quasi likelihood methods are in general not applicable if the count data are censored or truncated.

To illustrate the idea behind the semiparametric estimator for the Poisson mixture model, rewrite the marginal probability function (4.7) as

$$f(y) = \frac{\lambda^y}{y!} \mathrm{E}_u \left(e^{-\lambda u} u^y \right) \tag{4.25}$$

where E_u denotes the expectation with respect to the mixing distribution $g(u)$ which is left unspecified. Recall the definition of a moment generating function

$$M(s) = \int e^{sx} f(x) dx$$

Taking y-th order derivatives with respect to s, we get

$$\begin{aligned} M^{(y)}(s) &= \int e^{sx} x^y f(x) dx \\ &= \mathrm{E}(e^{sx} x^y) \end{aligned}$$

For $s = -\lambda$ and $x = u$, this is precisely the expectation on the right side of (4.25) so that we can write the Poisson-mixture probability function as

$$f(y|x) = \frac{\lambda^y}{y!} \cdot M_u^{(y)}(-\lambda) \tag{4.26}$$

where $M_u^{(y)}$ is the y-th order derivative of the moment generating function of u. The log-likelihood for a sample of n independent observations is

$$\ell = \sum_{i=1}^{n} \left[y_i \log \lambda_i - \log y_i! + \log M_u^{(y)}(-\lambda_i) \right] \tag{4.27}$$

Gurmu, Rilstone and Stern (1998) approximate $g(u)$ by Laguerre polynomials, derive the corresponding moment generating function, and use this function to estimate β together with additional parameters of the approximation by maximum likelihood, hence effectively avoiding the *a-priori* specification of a density function for the unobserved heterogeneity component. They show that the resulting estimator is consistent.

4.4.2 Finite Mixture Models

An alternative semiparametric approach to modeling unobserved heterogeneity has been popularized in econometrics by Heckman and Singer (1984).

Earlier references in the statistical literature include Simar (1976) and Laird (1978). The semiparametric maximum likelihood estimator is based on a finite mixture specification in which the underlying distribution of v (that is, of the intercept in the Poisson regression model) is approximated by a finite number of support points. This is a straightforward application of the mixture concept introduced in Chap. 2.5.1. The first application to the Poisson regression model with unobserved heterogeneity is due to Brännäs and Rosenqvist (1994).

In this approach, unobserved heterogneity is modeled as a discrete distribution with K classes and mass points $v \in \{v_1, \ldots, v_K\}$. If π_k denotes the probability $P(v = v_k)$, the marginal probability function for individual i is

$$f(y_i|x_i) = \sum_{k=1}^{K} \pi_k f_{ik}$$

where $f_{ik} = f(y_i|x_i, \beta, v_k)$ is the response distribution in the k-th component of the finite mixture, in this case a Poisson distribution with parameter

$$\lambda_{ik} = \exp(x_i'\beta + v_k)$$

where x_i does not include a constant term. For a given K, semi-parametric log likelihood function can be expressed as

$$\log L(\beta, v, \pi) = \sum_{i=1}^{n} \log \left\{ \sum_{k=1}^{K} \pi_k f_{ik} \right\} \tag{4.28}$$

Relative to the standard Poisson model, there are $2(K-1)$ additional unknown parameters. The number of classes, K, is unknown as well. Usually, the semi-parametric maximum likelihood estimation proceeds in two stages: first, K is treated as fixed, and (4.28) is maximized with respect to β, v, and π. Second, K is picked using formal model selection techniques.

EM algorithm

Direct maximization of (4.28) with respect to the full parameter vector is in general difficult, if not impossible. Instead, the EM algorithm has become the method of choice. Aitkin (1999) shows that

$$\frac{\partial \log L(\theta)}{\partial \theta} = \sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik} \frac{\partial \log f_{ik}}{\partial \theta} \tag{4.29}$$

where $\theta = (\beta, v_1, \ldots, v_K)$ and

$$w_{ik} = \frac{\pi_k f_{ik}}{\sum_{k=1}^{K} \pi_k f_{ik}} \tag{4.30}$$

represents the posterior probability that the i-th unit comes from the k-th component of the mixture. Solving these equations for a given set of weights,

and updating the weights from the current parameter estimates is an EM algorithm. For an alternative derivation, one can to think of this problem as a missing data problem. If true group membership was observed, we could write the complete data log-likelihood as

$$\log L(\theta) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} [\log(\pi_k) + \log f_{ik}] \tag{4.31}$$

where z_{ik} is a multinomial indicator variable of group membership. The E-step of the EM algorithm constructs the log-likelihood of *observed* data by taking expectations of the *complete* data log-likelihoods over the unobservable component indicators z_{ik}, conditional on y_i and x_i, and given parameter values for β and v in iteration round (r). But the conditional expectation of z_{ik}, evaluated at parameter values at step (r), $\hat{\beta}^{(r)}$ and $\hat{v}^{(r)}$, is just the weight $w_{ik}^{(r)}$ defined in (4.30). The estimated parameters in step (r) are then the solution of the following M-step equations:

$$\frac{\partial \log L(\theta)}{\partial \pi_k} = \sum_{i=1}^{n} \left\{ \frac{w_{ik}^{(r)}}{\pi_k} - \frac{w_{iK}^{(r)}}{\pi_K} \right\} = 0, \qquad k = 1, \ldots, K-1 \tag{4.32}$$

from where we find that the unconditional probability weights are simply the averages of the posterior probabilities from the previous step,

$$\hat{\pi}_k^{(r+1)} = \sum_{i=1}^{n} \frac{w_{ik}^{(r)}}{n} ,$$

and

$$\frac{\partial \log L(\theta)}{\partial \theta} = \sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik}^{(r)} \frac{\partial \log f_{ik}}{\partial \theta} = 0 \tag{4.33}$$

Since analytical solutions of (4.33) are unavailable for the Poisson regression model with unobserved heterogeneity, one can use a standard Newton-Raphson algorithm to obtain estimates. The E- and M-steps are alternated repeatedly, until the increase in log-likelihood from step (r) to step $(r+1)$ is arbitrarily small.

It is often remarked that the EM algorithm is sensitive to the starting values, as it may converge to local maxima. Presumably, this is less of a concern in the context of a Poisson regression model with unobserved heterogeneity, since the Poisson estimates are consistent and should provide reasonable starting values. Once the algorithm has converged for K components, new estimates are obtained for $K+1$ components. After that, a formal comparison of the K-class model and $K+1$-class model can be based on penalized likelihood criteria, such as AIC or BIC. Asymptotic standard errors can be obtained from the Hessian matrix of the log-likelihood function, evaluated at the parameter

estimates, by computing the square root of the diagonal elements of minus the inverted Hessian, as usual.

Mroz (1999) and Alfo and Trovato (2004) extend the discrete factor approach to multivariate data. In a related development, it is shown by Wedel et al. (1993) that this semiparametric estimator is readily extended to the case where heterogeneity not only affects the intercept but the regression coefficients as well (See also Wang et al., 1996). The model then takes the form

$$f(y|x) = \sum_{k=1}^{K} \pi_k \frac{\exp(-\exp(x'\beta_k))\exp(yx'\beta_k)}{y!} \tag{4.34}$$

where the intercept is part of x, with likelihood function given by

$$L(\beta, \pi) = \prod_{i=1}^{n} f(y_i; \beta_j) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k \frac{\exp(-\exp(x_i'\beta_k))\exp(y_i x_i'\beta_k)}{y_i!} \tag{4.35}$$

As Wedel et al. point out this model has an interesting interpretation in terms of sample segmentation. In particular, the estimated proportions $\hat{\pi}_1, \ldots, \hat{\pi}_K$ give the fraction of the sample that is of type 1, type 2 ... type K, respectively. The posterior probability is, as before,

$$w_{ik} = \frac{\pi f(y_i|x_i; \beta_k)}{\sum_{k=1}^{K} \pi_k f(y_i|x_i; \hat{\beta}_k)} \tag{4.36}$$

Heterogeneity, rather than being a nuisance factor that leads to overdispersion and complicates statistical inference, provides useful information on features of the population. In particular, a change in x_i may have a large effect in one group but no, or even a negative effect in another group. Wedel at al. (1993) discuss an application in marketing research where the dependent variable is the number of customer purchases of books offered through direct mail. A study of market segmentation shows that different groups of individuals react differently to various marketing techniques (such as mail offers, mail promotions, or sweepstakes). Moreover, the posterior probabilities can be used to determine the segment each customer falls in with highest probability. This ability of attributing individuals to market segments can thus provide valuable information for targeted marketing.

More recently, this finite mixture model has been extended in two directions. First, van Duijn and Böckenholt (1995) and Deb and Trivedi (1997) consider a finite mixture negative binomial model. Second, Wang, Cockburn, and Puterman (1998) parameterize the mixture probabilities as a function of further explanatory variables.

5

Sample Selection and Endogeneity

5.1 Censoring and Truncation

Unobserved heterogeneity can be interpreted as limited observability of *independent* variables. This section deals with limits in observability of the *dependent* variable which may stem from selective sampling or selective recording. The most common forms of sample selection arise from censoring and truncation. In the former case, the count dependent variable is observed only over a limited range, whereas in the latter case, certain observations are omitted entirely from the sample.

A common way to present such models is to distinguish between a latent count y^*, an observed count y, and a selection variable c. The binary variable c may indicate censoring, truncation, or non-reporting. We adopt the convention that truncation, censoring, or non-reporting occurs if $c = 0$. Models of "exogenous" censoring or truncation are based on the following mechanism:

$$c = \begin{cases} 1 \text{ if } y^* \in A \\ 0 \text{ if } y^* \notin A \end{cases} \tag{5.1}$$

that is, c is uniquely determined through the count dependent variable y^*. The two most commonly encountered situations are:

1. A is the set of positive integers (truncation/censoring at zero).
2. A is the set $\{0, \ldots, a\}$ where a is some positive integer (right truncation/censoring).

A common example of censored observations is given in Terza (1985). Here, the dependent variable is obtained from a survey question "How many times have you been to shopping area Q in the past thirty days?," with responses "zero", "one", "two", or "three or more". An example for truncated count data is the number of unemployment spells in the population of unemployed workers.

5.1.1 Truncated Count Data Models

The most common form of truncation is (left) truncation at zero. Truncated Poisson and negative binomial models have been discussed, among others, by Creel and Loomis (1990) and Grogger and Carson (1991). Gurmu (1991) refers to the truncated at zero Poisson model as "positive Poisson regression".

In the Poisson case, the observed data distribution is given by

$$\begin{aligned} f(y|x, y > 0) &= \frac{f(y, y > 0|x)}{f(y > 0|x)} \\ &= \frac{\exp(-\lambda)\lambda^y}{y!(1 - \exp(-\lambda))} \qquad y = 1, 2, \ldots \end{aligned}$$

where, as before, $\lambda = \exp(x'\beta)$. The truncated negative binomial model is obtained in a similar way. Mean and variance of the truncated-at-zero Poisson model are given by

$$\mathrm{E}_{\mathrm{tz}}(y|\lambda, y > 0) = \frac{\lambda}{1 - \exp(-\lambda)} \tag{5.2}$$

and

$$\mathrm{Var}_{\mathrm{tz}}(y|\lambda, y > 0) = E(y|\lambda, y > 0)\left(1 - \frac{\lambda}{\exp(\lambda) - 1}\right). \tag{5.3}$$

Since λ (the mean of the untruncated distribution) is greater than zero, $0 < \exp(-\lambda) < 1$ and the truncated mean is shifted to the right. Moreover, the truncated-at-zero model displays underdispersion since $0 < 1 - \lambda/(\exp(\lambda) - 1) < 1$.

Grogger and Carson (1991) apply the truncated model to the number of recreational fishing trips taken by a sample of Alaskan fishermen. Gurmu (1991) applies it to the Kennan (1995) data set on contract strikes. An important application of truncated-at-zero count data models is the use as a building block for hurdle models (see Chapter 6.3).

5.1.2 Endogenous Sampling

The sampling paradigm underlying truncated count data models is essentially one of random sampling: a draw is taken from the population. If that draw falls within the truncation area, the observation is dropped. Otherwise, it is kept. The probabilities of non-truncated outcomes are scaled upwards *proportionally*, relative to the population probabilities.

This approach may be inappropriate if the inclusion in the sample is conditioned on at least one occurrence. A typical example is a survey of shopping mall visitors, asking them for the number of shopping trips taken. Clearly, in such a case, the minimum response must be a count of one if the time frame

of the response includes the present time, and zeros are therefore unobserved. However, as first pointed out by Shaw (1988), such responses do not have the probability function of the truncated-at-zero count data model. The reason is that the inclusion in the sample is *endogenous* in the sense that more frequent users have a higher probability of being included in the sample than less frequent users. As a consequence, large counts are overrepresented in the sample relative to the population), and the distribution of positive counts in the sample is not proportional to the distribution of positive counts in the population. But the truncated models precisely requires such proportionality. Ignoring endogenous sampling leads to biased inference for the parameters of the population distribution, although truncated-at-zero models have been sometimes used for such samples (Grogger and Carson, 1991).

Shaw (1988) derives the likelihood function for an endogenous, or "on-site" sampling scheme. An alternative derivation is given by Santos Silva (1997a) who, in turn, applies results from Manski and Lerman (1977). To provide the intuition why high counts are overrepresented in an on-site sample, consider two individuals, the first with a count of one and the second with a count of two. The sample is taken "on-site" at a random point during the time interval. But then, an individual with a count of two is twice as likely to be included in the sample as an individual with a count of one. If $P(y = 1)$ and $P(y = 2)$ are the probabilities of outcomes one and two in the population, respectively, then the probabilities of outcomes one and two in the sample are given by $1 \times P(y = 1)$ and $2 \times P(y = 2)$, respectively. This reasoning can be generalized and, with appropriate normalization, we obtain

$$f^*(y|x) = \frac{f(y|x)y}{\sum_{k=1}^{\infty} f(k|x)k} = \frac{f(y|x)y}{\mathrm{E}(y|x)} \tag{5.4}$$

where $f^*(y|x)$ is the probability function of the sample. From here, it is clear that zeros are not observed and that individuals for which $y > \mathrm{E}(y|x)$ are overrepresented in the sample.

If the expressions for the Poisson probability function are substituted into (5.4) we obtain the on-site sample probability function

$$f_{\mathrm{os}}(y|\lambda) = \frac{\exp(-\lambda)\lambda^{y-1}}{(y-1)!}, \qquad y = 1, 2, 3 \ldots \tag{5.5}$$

Interestingly, this is exactly the probability function of a *shifted* or *displaced* Poisson distribution that is obtained by shifting the sample space by one count (Johnson and Kotz, 1969). The on-site sample Poisson distribution is also sometimes referred to as *size-biased Poisson*, or as Poisson distribtion with *endogenous stratification*.

Expected value and variance are given by

$$\mathrm{E}_{\mathrm{os}}(y|\lambda) = \sum_{k=0}^{\infty}(k+1)\frac{\exp(-\lambda)\lambda^k}{k!}$$

$$= \lambda + 1 \tag{5.6}$$

and

$$\begin{aligned} \mathrm{Var}_{\mathrm{os}}(y|\lambda) &= \sum_{k=0}^{\infty}(k+1)^2 \frac{\exp(-\lambda)\lambda^k}{k!} - (\lambda+1)^2 \\ &= \lambda\,, \end{aligned} \tag{5.7}$$

respectively. For λ close to zero, underdispersion is substantial but it vanishes asymptotically for $\lambda \to \infty$. Santos Silva (1997a) discusses estimation of the on-site sample Poisson model with unobserved heterogeneity. See also Santos Silva (2003) and the discussion in Chap. 6.3.5.

Englin and Shonkwiler (1995) derive the on-site sample probability function when the population has a negative binomial distribution. It is given by

$$f_{\mathrm{os,nb}}(y|\lambda,\alpha) = \frac{y\Gamma(y+\alpha)}{\Gamma(y+1)\Gamma(\alpha)}\alpha^{-y}\lambda^{y-1}[1+\lambda/\alpha]^{-(y+\alpha)} \tag{5.8}$$

with $\mathrm{E}(y|x) = \lambda + 1 + \alpha^{-1}\lambda$ and $\mathrm{Var}(y|x) = \lambda(1 + \alpha^{-1} + \alpha^{-1}\lambda + \alpha^{-2}\lambda)$. For $\alpha \to \infty$, these moments converge to those of the on-site Poisson model, as they should.

Other endogenous sampling models can be thought of. For instance, it is not necessarily the case that zeros are ruled out in such models. One example would be the case where inclusion in the sample depends on a "on-site" visit, but the question of interest refers to the number of visits in the preceding (rather than the current) month. Again, frequent users are over-represented in the data, but the required adjustment is different from (5.4).

5.1.3 Censored Count Data Models

Censored count data models have been studied by Terza (1985) and Brännäs (1992b), among others. Unlike for continuous data Tobit models, the type of censoring that is typically encountered in count data models is right-censoring. It arises in survey questionnaires where the highest category is "x or more" counts. The standard definition of a censored count data model is then based on the observation mechanism

$$c = \begin{cases} 1 \text{ for } y^* \in A = \{0,\ldots,a\} \\ 0 \text{ for } y^* \in A = \{a+1, a+2, \ldots\} \end{cases} \tag{5.9}$$

Thus, $P(c=1) = F(a)$ where $F = \sum_{j=0}^{a} f(j)$, and $P(c=0) = 1 - F(a)$. The probability function of observed counts y for individual i is equal to

$$g(y_i|x_i,c_i) = f(y_i)^{c_i}[1-F(a)]^{1-c_i} \tag{5.10}$$

and the log-likelihood function for a random sample of size n has now two components

$$\ell(\beta) = \sum_{i=1}^{n} c_i \log f(y_i) + (1 - c_i) \log(1 - F(a)) \tag{5.11}$$

The first term on the right gives the likelihood contribution of the non-censored observations, while the second term on the right gives the contribution of the censored observations. Terza (1985) provides details on implementing a Newton-Raphson algorithm for the maximum likelihood estimator $\hat{\beta}$.

Another type of censoring has been considered in Caudill and Mixon (1995). They are concerned with estimating the determinants of completed fertility using survey data of women, some of whom are still in their childbearing years, defined to be the age of forty or less. Let y^* denote completed fertility, measured by the (final) number of children in the family, and y denote current fertility. Then $y^* = y$ if age ≥ 40 and $y^* \geq y$ if age < 40. Define a variable

$$c = \begin{cases} 1 \text{ if age} \geq 40 \\ 0 \text{ if age} < 40 \end{cases} \tag{5.12}$$

The log-likelihood function can then be written as

$$\ell(\beta) = \sum_{i=1}^{n} c_i \log f(y_i) + (1 - c_i) \log(1 - F(y_i - 1)) \tag{5.13}$$

In contrast to the standard censoring model (5.11), the censoring threshold varies now from observation to observation. An interesting modification of this idea was recently introduced by McIntosh (1999). Rather than assuming that childbearing years end for each person at the age of 40, McIntosh derived the censoring status from an additional survey question on the desirability of further children (independently of age). In this case, the selection variable in (5.12) simply needs to be re-defined such that $c = 1$ if no further children are desired, and $c = 0$ else.

5.1.4 Grouped Poisson Regression Model

Closely related to censoring is the concept of grouping. The grouped Poisson regression model was discussed by Moffatt (1995) who showed that the resulting log-likelihood function is globally concave and the maximum likelihood estimator $\hat{\beta}$ thus unique (See also Moffatt, 1997b, and Moffatt and Peters, 2000). In particular, consider a mutually exclusive and exhaustive partition of $\mathbb{N}_0$ into J subsets $A_1, \ldots, A_J$. Also, assume that the subsets are ordered, such that each set consists of an uninterrupted sequence of natural numbers, and one plus the largest number in set A_j is equal to the smallest number in set A_{j+1}. Clearly,

$$p_{ij}(x_i) = P(y_i \in A_j | x_i) = \sum_{k \in A_j} f(k|x_i) \tag{5.14}$$

where $f(k)$ is the Poisson probability function. The log-likelihood function for a sample of independent observations can be written as

$$\ell(\beta) = \sum_{i=1}^{n} \sum_{j=1}^{J} d_{ij} \log p_{ij}(x_i) \tag{5.15}$$

where the variable d_{ij} indicates membership of individual i in group j. Maximization of the log-likelihood function can make use of the Newton-Raphson algorithm in the usual way.

5.2 Incidental Censoring and Truncation

So far, models in this chapter were based on a single population model. This underlying model at the same time described the population distribution as well as the censoring or truncation process. Separate determinants of outcome and observation mechanism were excluded, as were separate parameters for the two aspects of the model. The literature offers a number of suggestions for richer, and potentially more interesting models. Essentially, these generalizations allow for separate, although not necessarily unrelated, processes for the count outcome equation and for the selection mechanism.

5.2.1 Outcome and Selection Model

The canonical model for incidental censoring and truncation assumes that the population model of interest is a count data model with conditional expectation function

$$\mathrm{E}(y|x, v) = \exp(x'\beta + v) \tag{5.16}$$

where v represents unobserved heterogeneity as before. It would be hard to build a model of incidental censoring without introducing unobserved heterogeneity.

The selection equation is typically a binary response model with linear latent index function

$$c^* = z'\gamma + \varepsilon \tag{5.17}$$

that generates a binary indicator variable c such that

$$c = \begin{cases} 1 \text{ if } c^* \geq 0 \\ 0 \text{ if } c^* < 0 \end{cases} \tag{5.18}$$

Together, the two models for y and c can be combined as building blocks for models of incidental censoring or incidental truncation. In the former case, we observe a censored outcome (such as zero), rather than the original y, whenever $c = 1$. In the latter case, y is not observed at all whenever $c = 1$. This is the classical sample selection model, here for counts rather than for

a continuous dependent variable in a linear regression model, as in Heckman (1979). The conditional expectation in the selected sample is then

$$\mathrm{E}(y|x, c = 1) = \mathrm{E}_{v|x,c=1}[\exp(x'\beta + v)] = \exp(x'\beta)\mathrm{E}[\exp(v)|x, c = 1]$$

which is not equal to $\exp(x'\beta)$ in general even though $\mathrm{E}[\exp(v)|x] = 1$. As a consequence, estimating a count data model in the selected sample leads to biased estimates of the population parameters. There is one important exception when the CEF in the selected sample is equal to the population CEF, namely when v and c are independent, which is the same as saying that v and ε are independent. Thus, special models for incidental censoring and truncation are only required to the extent that v and ε are correlated. There are typically many good reasons for suspecting such a correlation, depending on the particular application. Generally speaking, correlation must arise whenever there are common unobserved factors that affect both outcome and selection equation.

5.2.2 Models of Non-Random Selection

This brings us to a crucial question, namely how the joint distribution of, and thus any correlation between, v and ε should be modeled. The leading assumption in the literature is that v and ε have a bivariate normal distribution (see Terza, 1998, Winkelmann, 1998, Deb and Trivedi, 2006, and the related literature for linear models). In this approach, dependence is described by a single parameter, ρ, which can vary between -1 and +1, with the value of zero corresponding to independence. Maximum likelihood estimation of such models is relatively straightforward, although it requires the numerical evaluation of a one-dimensional integral. Note that the bivariate normal assumption implies that the marginal distribution of the count dependent variable under independence is of a Poisson-log-normal form, an assumption that is quite appealing, as it appears dominate the more conventional negative binomial model in many empirical applications (see the remarks in the previous chapter).

On the downside, the bivariate normal model imposes restrictions that may or may not hold in a particular application. While it can therefore serve as a natural starting point for modeling incidental truncation and censoring, it is natural that alternative estimation methods have been pursued in the literature.

One such alternative, proposed by Weiss (1999) and building on Lee (1983), is a transformation approach. For example, assume that $\exp(v)$ has a gamma distribution so that the marginal distribution of $y|x$ is negative binomial. Consider the transformation

$$h(v) = \Phi^{-1}[G(\exp(v))]$$

where Φ is the cumulative density function of the standard normal distribution and G is the cumulative density function of the gamma distribution. As a result, $h(v)$ has a standard normal distribution. Moreover, Weiss (1999) lets

$$\varepsilon = \rho h(v) + \eta$$

where η is independently normally distributed with mean zero and variance σ_η^2. Thus, $h(v)$ and ε are bivariate normal, although v is not.

This approach is a special case of the more general copula approach for non-normal data (Van Ophem, 2000, and Zimmer and Trivedi, 2006). Copula based models can in general be estimated without resorting to numerical integration or simulation. A drawback is that copulas place restrictions on the pattern of allowable correlations.

It is also possible to specify the error structure using discrete distributions as described by Mroz (1999). Because such models are finite mixture models, they are semiparametric and the discrete distributions can, in principle, approximate any continuous distributions. However, Deb and Trivedi (2006) report serious implementation issues and convergence problems, estimates from alternative runs that gave effects of different signs, significance and magnitude, indicating multiple maxima. Thus, while theoretically appealing, this approach may be less so from a practical perspective. Finally, semi-parametric count selection models based on series expansions are discussed in Romeu and Vera-Hernandez (2005) whereas Lee (2004) considers certain non-parametric approaches.

5.2.3 Bivariate Normal Error Distribution

In the canonical selection model, v and ε are bivariate normally distributed with mean vector zero and covariance matrix

$$\Sigma = \begin{bmatrix} \sigma_v^2 & \sigma\rho \\ \sigma\rho & 1 \end{bmatrix} \tag{5.19}$$

where ρ is the coefficient of correlation between v and ε, and σ_v^2 is the variance of v. The variance of ε is normalized to one, since it is not identified in the probit selection equation.

The following derivations make use of results for conditional distributions involving two jointly normally distributed random variables (here: ε and v). In particular, we are interested in the following three conditional distributions, as well as their first moments

a) $f(\varepsilon|v)$
b) $f(v|\varepsilon > -z'\gamma) = f(v|c = 1)$
b) $f(\exp(v)|\varepsilon > -z'\gamma) = f(\exp(v)|c = 1)$

A general result related to a) is that if x_1 and x_2 are bivariate normal with means μ_1 and μ_2, variances σ_1^2 and σ_2^2, and correlation ρ, then

$$x_2|x_1 \sim N\left[\mu_2 + \rho\sigma_2\left(\frac{x_1 - \mu_1}{\sigma_1}\right), \sigma_2^2(1 - \rho^2)\right] \tag{5.20}$$

Thus, under the assumptions on the mean and variances of ε and v, it follows that

$$\varepsilon|v \sim N\left[\rho v/\sigma_v, 1-\rho^2\right] \tag{5.21}$$

Similarly,

$$v|\varepsilon \sim N\left[\rho\sigma\varepsilon, \sigma_v^2(1-\rho^2)\right] \tag{5.22}$$

Moreover, the required distribution for b) can be written in general terms as

$$f(v|\varepsilon > -z'\gamma) = \int_{-z'\gamma}^{\infty} \frac{\phi_2(v,\varepsilon)}{1-\Phi(-z'\gamma)} \mathrm{d}\varepsilon$$

where ϕ_2 is the bivariate standard normal density function, and Φ is the cumulative density function of the univariate standard normal distribution. No closed form solution is available. To derive the conditional expectation of the incidentally censored error v, note first that in the univariate normal case

$$\mathrm{E}(\varepsilon|\varepsilon > -z'\gamma) = \frac{\phi(-z'\gamma)}{1-\Phi(-z'\gamma)}$$

(see Maddala, 1983). Moreover, from (5.20), $\mathrm{E}(v|\varepsilon) = \rho\sigma\varepsilon$. Thus,

$$\begin{aligned} \mathrm{E}(v|\varepsilon > -z'\gamma) &= \rho\sigma \mathrm{E}(\varepsilon|\varepsilon > -z'\gamma) \\ &= \rho\sigma \frac{\phi(-z'\gamma)}{1-\Phi(-z'\gamma)} \\ &= \rho\sigma \frac{\phi(z'\gamma)}{\Phi(z'\gamma)} \end{aligned} \tag{5.23}$$

Similarly

$$\mathrm{E}(v|\varepsilon < -z'\gamma) = -\rho\sigma \frac{\phi(z'\gamma)}{1-\Phi(z'\gamma)} \tag{5.24}$$

The derivation of the last quantity of interest,

$$\mathrm{E}[\exp(v)|\varepsilon > -z'\gamma]\,,$$

is somewhat more complex. We know in general that if $v|\varepsilon$ is normally distributed (see equation(5.22)), then $\exp(v)|\varepsilon$ must have a log-normal distribution. Therefore, an expression such as $\mathrm{E}[\exp(v)|\varepsilon > -z'\gamma]$ is the mean of a log-normal distribution with incidental censoring.

In a first step, it is relatively straightforward to derive $\mathrm{E}(\exp(v)|\varepsilon)$. We know that $v|\varepsilon$ has a normal distribution with mean $\mathrm{E}(v|\varepsilon) = \rho\sigma_v\varepsilon$ and variance $\mathrm{Var}(v|\varepsilon) = \sigma_v^2(1-\rho^2)$. The moment generating function of a normally distributed random variable x is :

$$M_x(t) = \mathrm{E}(\exp(tx)) = \exp\left(\mu t + \frac{\sigma_v^2 t^2}{2}\right) \tag{5.25}$$

Therefore

$$\mathrm{E}(\exp(v)|\varepsilon) = M_{v|\varepsilon}(1) = \exp\left(\rho\sigma_v\varepsilon + \frac{\sigma_v^2(1-\rho^2)}{2}\right) \tag{5.26}$$

We ultimately are interested in $\mathrm{E}(\exp(v)|\varepsilon > -z'\gamma)$ which is equal to

$$\begin{aligned}\mathrm{E}(\exp(v)|\varepsilon > -z'\gamma) &= \int_{-z'\gamma}^{\infty} \mathrm{E}(\exp(v)|\varepsilon) f(\varepsilon|\varepsilon > -z'\gamma) d\varepsilon \qquad (5.27)\\ &= \exp\left(\frac{\sigma_v^2(1-\rho^2)}{2}\right) \times \frac{\int_{-z'\gamma}^{\infty} \exp\left(\rho\sigma_v\varepsilon - \frac{1}{2}\varepsilon^2\right) d\varepsilon}{\sqrt{2\pi}\Phi(z'\gamma)}\end{aligned}$$

The numerator of the second term on the right can be rewritten as

$$\int_{-z'\gamma}^{\infty} \exp\left(-\frac{1}{2}(\varepsilon - \rho\sigma_v)^2 + \frac{1}{2}\rho^2\sigma_v^2\right) d\varepsilon = \exp\left(\frac{1}{2}\rho^2\sigma_v^2\right)\sqrt{2\pi}\Phi(z'\gamma + \rho\sigma_v)$$

and therefore (see also Terza, 1998, and Johnson, Kotz and Balakrishnan, 1994)

$$\mathrm{E}(\exp(v)|\varepsilon > -z'\gamma) = \exp(\sigma_v^2/2)\frac{\Phi(z'\gamma + \rho\sigma_v)}{\Phi(z'\gamma)} \tag{5.28}$$

This conditional expectation simplifies to the expectation of a log-normal distribution for $\rho = 0$, as it should.

The Poisson-bivariate-normal set-up allows for incidental selection in the sense that unobserved factors affecting c also affect y^*. Ignoring this correlation will lead to a misspecified model with the possibility of inconsistent parameter estimates. The effects of selectivity in count data models are similar to those found for continuous data (See e.g. Heckman, 1979), and the corresponding models will be presented in the next sections. Existing models for incidental truncation and censoring are presented there. Other models with similar structure (endogenous switching and endogenous reporting) follow in a later part of this chapter.

5.2.4 Outcome Distribution

Before proceeding, it is instructive to consider the implications of alternative distributional choices for y in this class of models. It has been emphasized earlier that models can be specified with either Poisson or negative binomial distribution. By contrast, the negative binomial distribution is no longer suitable in the context of endogenous selectivity. The resulting model suffers from overparameterization. To illustrate this point, consider the Negbin II model with $\mathrm{E}(y|x,u) = \lambda u$, where $\lambda = \exp(x'\beta)$, $u = \exp(v)$ and $\mathrm{E}(u|x) = 1$. Conditional on u, this model has the standard variance function of the Negbin II model:

$$\mathrm{Var}(y|x,u) = \lambda u + \alpha(\lambda u)^2$$

where α is the dispersion parameter. Marginalizing with respect to u yields

$$\mathrm{E}(y|x) = \mathrm{E}_u[\mathrm{E}(y|u)] = \lambda \tag{5.29}$$

and

$$\begin{aligned}\mathrm{Var}(y|x) &= \mathrm{E}_u[\mathrm{Var}(y|x,u)] + \mathrm{Var}_u[\mathrm{E}(y|x,u)] \\ &= \lambda + \lambda^2(\sigma_u^2 + \alpha\sigma_u^2 + \alpha) \\ &= \lambda + c\lambda^2\end{aligned} \tag{5.30}$$

Let $\sigma_u^2 = (c-\alpha)/(1+\alpha)$ and it is clear that α and σ_u^2 are not identified in the variance function. Identification therefore depends on higher order moments. This is unsatisfactory and may lead to numerical instability in practice.

5.2.5 Incidental Censoring

A model for incidental censoring was introduced by Crépon and Duguet (1997b). In this model, which is formally equivalent to the class of zero-inflated count data models to be discussed in Chap. 6.4,

$$y = \begin{cases} 0 & \text{if } c = 1 \\ y^* & \text{if } c = 0 \end{cases} \tag{5.31}$$

and $c = 1$ for $\varepsilon > -z'\gamma$. Recall from Chap. 5.2.3 that the conditional distribution of ε given v is a normal distribution with mean $\rho v/\sigma$ and variance $1-\rho^2$. As a consequence, we can express the probability of censoring ($c = 1$) *conditional* on the error in the outcome equation as

$$\begin{aligned}P(c=1|v) &= P(\varepsilon > -z'\gamma|v) \\ &= 1 - \Phi\left(\frac{-z'\gamma - \rho v/\sigma}{\sqrt{1-\rho^2}}\right) \\ &= \Phi\left(\frac{z'\gamma + \rho v/\sigma}{\sqrt{1-\rho^2}}\right) \\ &\equiv \Phi^*(v)\end{aligned} \tag{5.32}$$

where Φ is the cumulative density function of the standard normal distribution. Furthermore, for $y^*|v \sim$Poisson with $\tilde{\lambda} = \exp(x'\beta + v)$, the probability function of y, conditional on v, is given by

$$f(y|v,x,z) = \Phi^*(v)d + [1 - \Phi^*(v)]\frac{\exp(-\tilde{\lambda})\tilde{\lambda}^y}{y!} \tag{5.33}$$

where $d = 1 - \min\{y, 1\}$. This probability function depends on the unobserved v, and the observed data distribution (not conditioned on v) can be obtained after integrating the joint distribution of y and v over v:

$$f(y|x,z) = \int_{-\infty}^{\infty} f(y|v,x,z)f(v)\mathrm{d}v \tag{5.34}$$

$$= \int_{-\infty}^{\infty} \left\{ \Phi^*(v)d + [1 - \Phi^*(v)] \frac{\exp(-\tilde{\lambda})\tilde{\lambda}^y}{y!} \right\} f(v)\mathrm{d}v$$

where $f(v)$ is the density function of the marginal distribution of v, in this case a normal distribution with mean 0 and variance σ^2. While marginalizing with respect to v does not lead to a closed form solution, Crépon and Duguet suggest a feasible simulation method due to Gourieroux and Monfort (1993). Alternatively, Gauss-Hermite quadrature can be used for an approximate evaluation of the integral (See Appendix B). Crépon and Duguet apply their model to a study of R&D productivity, where y^* gives the number of discoveries, y the number of patents applied for, and $c = 1$ if the firm decided to apply for patents in general.

5.2.6 Incidental Truncation

A model for endogenous truncation based on the bivariate normal distribution has been studied, among others, by Greene (1998) and Winkelmann (1998). This model is the count data version of the classical sample selection model (Heckman, 1979). It can be written as

$$y = \begin{cases} y^* & \text{if } c = 1 \\ \text{unobserved} & \text{if } c = 0 \end{cases} \tag{5.35}$$

where c and y^* are determined as above, i.e., $y^* \sim \text{Poisson}(\tilde{\lambda})$ with $\tilde{\lambda} = \exp(x'\beta)u$, $c = 1$ for $\varepsilon > -z'\gamma$, and ε and $\log(u)$ are bivariate normal distributed with correlation ρ.

What are the consequences of endogenous selectivity when this process is ignored and a standard count data model is estimated on the selected subsample of data? In general, there will be a bias both in the estimated constant and in the estimated slope parameters. Under some specific (and highly unrealistic) assumptions, the bias may be limited to the constant. To understand why such sample selection can not be simply ignored, recall a critical identifying assumption in the standard model for unobserved heterogeneity, namely that u, and therefore $v = \log(u)$, is independently distributed of x. This assumption is likely to be violated in the sample selection model.

In fact, the nature of the inconsistency can be pinned down more precisely by considering a result from Chap. 5.2.3 on the expected value of an incidentally truncated log-normal variable. We can write for the sample of observed counts

$$\begin{aligned} \mathrm{E}(y|x, c = 1) &= \mathrm{E}_{u|c=1}\mathrm{E}(y|x, u, c = 1) \\ &= \exp(x'\beta)\mathrm{E}(u|x, c = 1) \end{aligned} \tag{5.36}$$

where $u = \exp(v)$ as before. Using (5.28), we obtain

$$\mathrm{E}(u|x, c = 1) = \exp(\sigma_v^2/2) \frac{\Phi(z'\gamma + \rho\sigma_v)}{\Phi(z'\gamma)} \tag{5.37}$$

and therefore

$$\begin{aligned}\mathrm{E}(y|c=1) &= \exp(x'\beta + \sigma_v^2/2)\frac{\Phi(z'\gamma + \rho\sigma_v)}{\Phi(z'\gamma)} \\ &= \exp(x'\beta^*)Q(\theta,\gamma,z) \end{aligned} \tag{5.38}$$

where β^* is the same as β except that the intercept is shifted by $\sigma_v^2/2$, and $\theta = \rho\sigma_v$. We see that $Q(\theta,\gamma,z) = 1$ if $\rho = 0$, so there is no problem in this case. If $\rho \neq 0$, $\mathrm{E}(y|c=1,x) \neq \exp(x'\beta^*)$. In particular, unless x and z are fully independent (which, at a minimum, rules out any overlap between the two sets of explanatory variables), the conditional expectation function in the selected sample is not proportional to $\exp(x'\beta)$, which means that the maximum likelihood estimator in the selected model is inconsistent.

Estimation

There are in principle two methods to obtain consistent parameter estimates in this framework. The first exploits the full parametric structure of the model and estimates all model parameters jointly by full information maximum likelihood. Estimation is then based on the probability function, for individual i,

$$\begin{aligned} f(y_i, c_i = 1) &= \int_{-\infty}^{\infty} f(y_i, c_i = 1|v_i) f(v_i) \mathrm{d}v_i \\ &= \int_{-\infty}^{\infty} f(y_i|v_i) \Phi^*(v_i) f(v_i) \mathrm{d}v_i \end{aligned} \tag{5.39}$$

where we applied the same factorization as before, i.e., $f(y_i|v_i)$ is a Poisson distribution, Φ^* is defined as in (5.32), and Gauss-Hermite quadrature is required.

Second, one can forego efficiency and base estimation on the first-order moment only. The conditional expectation of the observed sub-sample was given in (5.38). This model can be estimated directly by non-linear least squares.

Alternatively, a first order Taylor-series expansion of $\log Q(\theta)$ around $\theta = 0$ yields (Greene, 1998)

$$\begin{aligned} \log Q &\approx \rho\sigma \frac{\phi(z'\gamma)}{\Phi(z'\gamma)} \\ &= \rho\sigma m \end{aligned} \tag{5.40}$$

Thus, two-step estimation of a Poisson model with endogenous truncation can be justifiably based on a conditional expectation function

$$\lambda^* = \exp(x'\beta + \tau\hat{m}) \tag{5.41}$$

which can be estimated by a two-step procedure: a probit regression provides a consistent estimator $\hat{\gamma}$. The predicted selectivity term $\hat{m} = \phi(z'\hat{\gamma})/\Phi(z'\hat{\gamma})$ is then used as a regressor in a second step Poisson regression. The asymptotic

covariance matrix for $\hat{\beta}$ and $\hat{\tau}$ must be adjusted for the fact that the estimated inverse Mills ratio is a generated regressor (Murphy and Topel, 1985). Without correction, the estimated standard errors provide lower bounds for the true standard errors.

This approach was used by Freund, Kniesner and LoSasso (1999) in a study of health care utilization where the goal was to correct for potentially endogenous sample attrition. Greene (1998) had to correct for selective credit card approval in a study of the individual determinants of credit card default.

If $\rho \neq 0$, it is difficult to assess how reasonable the approximation is, and bias may be substantial. By contrast, this approximation can be very useful to obtain a simple test for exogeneity. Under the null hypothesis $H_0 : \rho = 0$, maximum likelihood estimation including the generated inverse Mills ratio is consistent and the distribution of the Wald test statistic is well-defined. For a test, it is much easier to estimate a first-stage probit and a second stage Poisson quasi-likelihood model than it is to implement a full information maximum likelihood estimation based on (5.39).

5.3 Endogeneity in Count Data Models

5.3.1 Introduction and Examples

Endogeneity describes a situation where inference on the structural relationship between two or more variables can not be based simply on a conditional model (conditional distribution function or conditional expectation function). Endogeneity is the absence of exogeneity. We will discuss two formal definitions of exogeneity below, one based on parameter ancillarity, and one based on a ceteris-paribus interpretation of the log-linear regression model embedded in (most) count data models. In most cases, the two approaches amount to the same thing, and as for the linear model, endogeneity essentially arises due to dependence between explanatory variables and the stochastic error term. Dependence may arise due to omitted variables that are correlated with the included ones, or, more generally, due to a simultaneous determination of the explanatory variables through a related model.

An important example where the issue of endogeneity is a major worry is related to the effect of a (binary) treatment on a count outcome variable. In experimental sciences, individuals are randomly assigned to treatment group and control group, and differences in outcomes will thus be a good estimator of the treatment effect. An example is the effect of a drug on the number of epileptic seizures (Diggle, Liang and Zeger, 1995).

In observational data, treatment is not assigned randomly. In many instances, individuals self-select into treatment, i.e., treatment becomes a matter of choice. For example, the number of doctor consultations may depend on the health insurance status (the "treatment" variable in this case). But insurance coverage is a choice variable that can depend, among other things,

on health status and the expected number of doctor visits itself. A moral hazard argument would suggest that health insurance might have a direct effect on the demand for health services. However, the adverse selection argument suggests that any observed correlation does not need to measure moral hazard per se. It could be the case that individuals with a latent high demand for health services are the ones who purchase more insurance. This is an instance of endogeneity due to an omitted variable, the "true health status". Even with very good data on health status (rarely available in practice) there can always be some residual uncertainty that would then cause a violation of the exogeneity assumption.

A similar problem is encountered if one wants to measure the effect of regular exercise on a person's health. Plausibly, healthier people are more likely to exercise regularly, which in turn may contribute to their good health. Disentangling these effects is a challenging task. Note that in both examples, the potentially endogenous x variable could be measured on a number of different scales:

- binary (health insurance yes/no; sport yes/no)
- ordinal (Plan A, B, and C ordered by level of generosity; sport never, at least once per month, at least once per week...).
- multinomial (private insurance, public insurance, no insurance; aerobic, anaerobic exercise).
- quantitative (out of pocket expense in percent of all expenses; minutes and intensity of exercise).

Endogeneity may also arise outside of the usual treatment effect framework. For instance, Mullahy (1997a) has studied how past cigarette consumption affects present cigarette consumption in order to assess the empirical evidence for addiction. Clearly, estimating addiction effects from observational data is a tall order, as unobserved smoking preferences will affect both past and present consumption. Nevertheless, methods are available to identify addiction effects in this set-up, as well as the treatment affects in the above examples. Those methods are the subject of this chapter.

5.3.2 Parameter Ancillarity

The first formal definition of exogeneity rests on parameter ancillarity. We are interested in the relationship between y_1 and y_2, where y_1 is a count and y_2 may or may not be a count. Since, as argued, count data should be modeled by way of conditional probability functions, a natural point to start the discussion of endogeneity is the joint distribution of y_1 and y_2, conditional on x:

$$f(y_1, y_2|x) = f(y_1, y_2|x, \theta_1, \theta_2) \tag{5.42}$$

where θ_1 and θ_2 are structural parameters. Once the full joint model is known and specified, it is always possible to use the maximum likelihood approach

to estimate all parameters of interest by full information maximum likelihood. However, it may sometimes be both impractical, and also unnecessary, to specify the full model. Suppose we want to estimate θ_1 only. Any joint distribution can be factored into a conditional and a marginal distribution. According the Engle, Hendry and Richard (1983), y_2 is called exogenous if we can base inference for θ_1 on the conditional model $f(y_1|y_2, x)$ alone, i.e., write

$$f(y_1, y_2|x; \theta_1, \theta_2) = f(y_1|y_2, x; \theta_1) f(y_2|x; \theta_2) \tag{5.43}$$

where θ_2 is an ancillary parameter. In this case, there is no loss of information when inference on θ_1 is based on the conditional model.

Example

Consider a simple bivariate linear model with feedback:

$$y_1 = \alpha y_2 + u_1$$

$$y_2 = \beta y_1 + u_2$$

If u_j are independently normal distributed with mean zero and variance σ_j^2, it follows that

$$f(y_1, y_2; \alpha, \beta, \sigma_1^2, \sigma_2^2) = f(y_1|y_2; \alpha, \sigma_1^2) f(y_2; \alpha, \beta, \sigma_1^2, \sigma_2^2)$$

where the marginal distribution of y_2 (the "reduced form") is a normal distribution with mean zero and variance $(\beta^2\sigma_1^2 + \sigma_2^2)/(1 - \beta\alpha)^2$. The parameters of the conditional model, α and σ_1^2, appear in the marginal distribution of y_2. Therefore, based on the Engle, Hendry and Richard (1983) definition, y_2 is not exogenous, and therefore endogenous.

The example allowed for a simple verification of the exogeneity condition, because the normal distribution has simple (normal) expressions for all three – joint, conditional, and marginal – distributions. Count data distributions do not have this property. In general, marginals and conditionals do not belong to the same family of distributions. To give an example, if $y_1|y_2$ is Poisson distributed, then the marginal distribution of y_1 cannot be a Poisson distribution. From the variance decomposition theorem, it always is the case that

$$\mathrm{Var}(y_1) = \mathrm{Var}[\mathrm{E}(y_1|y_2)] + \mathrm{E}[\mathrm{Var}(y_1|y_2)]$$

But if the conditional distribution is Poisson, we know that $\mathrm{Var}(y_1|y_2) = \mathrm{E}(y_1|y_2)$, and therefore $\mathrm{E}[\mathrm{Var}(y_1|y_2)] = \mathrm{E}(y_1)$, which shows that $\mathrm{Var}(y_1) > \mathrm{E}(y_1)$. Thus, there must be overdispersion at the marginal model, and hence $f(y_1)$ cannot be Poisson distributed.

5.3.3 Endogeneity and Mean Function

An alternative definition of exogeneity is not based on the full joint distribution of variables but rather centers on conditional expectations. The standard approach for endogeneity starts from a model with multiplicative unobserved heterogeneity, as in (4.1), whereby

$$\mathrm{E}(y|x,u) = \exp(x'\beta + \log(u)) = \exp(x'\beta)u$$

Endogeneity arises whenever $\mathrm{E}(u|x)$ is a function of x, rather then a constant (for convenience of notation normalized to 1), which implies that

$$\mathrm{E}(y|x) \neq \exp(x'\beta)$$

As a consequence, standard count data models, such as Poisson or negative binomial regression, do not identify β, the parameter of interest. This is the essential idea of endogeneity in count data models. Exogeneity, on the other hand, means that

$$\mathrm{E}(y|x) = \exp(x'\beta) \tag{5.44}$$

or, equivalently,

$$\mathrm{E}(u|x) = \mathrm{E}(y\exp(-x'\beta)|x) = 1 \tag{5.45}$$

Regressors are exogenous if the multiplicative stochastic error u is (mean) independent of x. The regressors are endogenous if $\mathrm{E}(u|x) \neq 1$.

Other, related formulations of the problem are possible. For example, we can always re-write

$$\begin{aligned}\mathrm{E}(y|x,u) &= \exp(x'\beta)u \\ &= \exp(x'\beta) + \exp(x'\beta)(u-1)\end{aligned}$$

For $\eta = \exp(x'\beta)(u-1)$ we obtain an additive model such that

$$\mathrm{E}(y|x,\eta) = \exp(x'\beta) + \eta \tag{5.46}$$

Under exogeneity, the multiplicative and the additive formulations are equivalent, since

$$\mathrm{E}(\eta|x) = \exp(x'\beta)(\mathrm{E}(u|x) - 1)$$

and $\mathrm{E}(\eta|x) = 0$ if and only if $\mathrm{E}(u|x) = 1$. The multiplicative interpretation is usually preferred, as it treats observed and unobserved heterogeneity symmetrically.

For yet another formulation, one could have started from a general additive model for y:

$$y = \exp(x'\beta) + \nu \tag{5.47}$$

It follows that

$$\mathrm{E}(y|x) = \exp(x'\beta) + \mathrm{E}(\nu|x)$$

which is the same as the expectation of (5.46) over η, with ν taking the role of η. However, (5.47) is an awkward expression to work with if y is a count variable, where we better think in terms of a probability model. For example, the decomposition (5.47) tells us that *if* ν is kept constant, it is the case that

$$\frac{\partial y/y}{\partial x} = \beta$$

But y takes on only integer values, so it is hard to interpret the above relative partial effect. The analogy to the linear model apparently stops here, and we will concentrate on conditional expectations for most of the remainder of this chapter.

In summary, we can state three equivalent definitions of exogeneity, $\mathrm{E}(y|x) = \exp(x'\beta)$, $\mathrm{E}(\eta|x) = 0$, or $\mathrm{E}(u|x) = 1$. Any of these is a key condition for consistency of the Poisson maximum likelihood or pseudo maximum likelihood estimator. A violation leads to an inconsistent Poisson PML, and alternative methods are required.

Possible Remedies

The remedies depend on the particular model structure. However, they can all be subsumed in one of five broad empirical strategies that we briefly outline first. More details are provided in the following sections.

The first approach is to ignore the count structure of the data and to approximate the conditional expectation function by a linear regression. This approach is advocated as an alternative to parametric limited dependent variable models by Angrist (2001). A similar argument can be made for count data. Once one settles for the linear model, the well-established tools for dealing with endogeneity in linear models can be applied. The downside is that the set of possible inferences is rather limited. No statements of the outcome distribution, or the data generating process, can be made. This method may work satisfactorily if all counts are large. Otherwise, the approximation error may be large.

A second approach uses a non-linear instrumental variable technique. This approach can be based either on the multiplicative or the additive model. Mullahy (1997a) recommends to use the moment condition

$$\mathrm{E}(u|z) = 1 \text{ , i.e., } \mathrm{E}\left(\frac{y}{\exp(x'\beta)} - 1 \middle| z \right) = 0 \tag{5.48}$$

where z are the instruments. Order conditions apply. Windmeijer and Santos Silva (1997) point out that instruments valid for a multiplicative error u are not necessarily vaild for an additive error η.

An alternative instrumentation can be implemented by estimation in stages. The assumptions are stronger than those required for IV estimation. In particular, assume that a single endogenous regressor has a reduced form equation

$$x = z'\gamma + \varepsilon \tag{5.49}$$

and that the instruments z are *fully independent* of ε and u. Then it is possible to estimate the effect of x on y in a standard count data regression, where x is replaced by $\hat{x} = z'\hat{\gamma}$ and $\hat{\gamma}$ is obtained from a first-stage estimation of model (5.49).

A fourth approach to deal with endogenous regressors is available if repeated measurements on the dependent and independent variables, say over time, are available and if endogeneity is caused by time invariant correlated unobserved heterogeneity. In this case, one can estimate the model parameters consistently by fixed effects panel data methods. Such methods are discussed in Chapter 7.

Finally, one can make parametric assumptions and specify the full joint distribution of y and x. For example, if x is binary, this approach leads to a Roy (1951) type switching regression model for count data.

5.3.4 A Two-Equation Framework

In Chap. 5.3.2, we considered endogeneity in a system of two linear equations with normally distributed errors. The properties of such a system are well understood (for instance identification through exclusion restrictions) and it appears worthwhile to study the extent to which the analogy of the linear model carries over to count data models. Such an analysis has been undertaken by Windmeijer and Santos Silva (1997) who consider the following system of equations:

$$y_1 = \exp(\alpha y_2 + x'\beta) + \nu \tag{5.50}$$

$$y_2 = \delta y_1 + z'\gamma + \varepsilon \tag{5.51}$$

where y_1 is a count dependent variable and y_2 is a second variable with unspecified scale.

As innocuous as the exponential function in the first equation may look, its consequences are far-reaching, as there are no well-defined reduced form equations. We can substitute and re-write, for instance, the first equation as

$$y_1 = \exp[\alpha(\delta y_1 + z'\gamma + \varepsilon) + x'\beta] + \nu$$

but we cannot solve for y_1 in general. The same holds for the second equation. Without the ability of converting one of the equations into a reduced form (or marginal) model, the system of two conditional models cannot be re-written as a joint distribution model, and identification of the structural parameters is infeasible.

A more optimistic conclusion is obtained once we introduce one additional restriction, either $\alpha = 0$ or $\delta = 0$. The system is then recursive, and reduced forms are well defined. For instance (the more interesting case), with $\delta = 0$, we obtain

$$\begin{aligned} y_1 &= \exp[\alpha(z'\gamma + \varepsilon) + x'\beta] + \nu \\ &= \exp[\alpha z'\gamma + x'\beta]\exp(\alpha\varepsilon) + \nu \end{aligned} \tag{5.52}$$

which identifies the parameters of interest in principle (except for the constant of the model - but this is not needed to estimate proportional effects), as long as ν and ε are independent of x and z, regardless of whether ν and ε are correlated or not.

Binary Endogenous Variable

A separate issue arises if y_2 is a binary variable. In this case, the second equation (5.51) of the two equation system has the interpretation of a latent model and the observed binary variable is determined by the threshold process

$$y_2 = \mathbb{1}(\delta y_1 + z'\gamma + \varepsilon > 0) \tag{5.53}$$

where $\mathbb{1}$ is the usual indicator function. In this case, logical consistency requires that for the unconditional probabilities, $P(y_2 = 1) + P(y_2 = 0) = 1$ which can be shown to imply that the system must be recursive. For example, for $x = z = 0$, y_2 determined as in (5.53) and $y_1 = \exp(\alpha y_2) + \nu$, we obtain

$$P(y_2 = 1) = P(\delta(\exp(\alpha \times 1) + \nu) + \varepsilon > 0)$$

$$P(y_2 = 0) = P(\delta(\exp(\alpha \times 0) + \nu) + \varepsilon \leq 0)$$

and thus

$$1 - F(\delta(-\exp(\alpha)) + F(-\delta) = 1$$

where F is the cumulative density function of $\delta\nu + \varepsilon$. Thus, it must be the case that either $\delta = 0$ or $\alpha = 0$ (or both). From a count data perspective, excluding the count from the binary equation is the more interesting model, yielding a count model with binary endogenous variable.

5.3.5 Instrumental Variable Estimation

A count data model with endogenous regressors can be estimated using instrumental variables, provided, of course, that instruments are available. Due to the exponential conditional expectation function, closed form solutions are unavailable, and one needs to apply non-linear instrumental variables techniques. A general exposition of the method is provided by Mullahy (1997a), who approached the issue in the context of the Poisson regression model. Windmeijer and Santos Silva (1997) discuss GMM estimation of such models. Further applications to count data modeling with endogeneity using instruments and GMM estimation include Vera-Hernandez (1999) and Schellhorn (2001).

Since the starting point is the exponential conditional expectation function rather than the full distribution, the method immediately generalizes to any

exponential regression model. By the same token, this also means that the technique does not generalize to arbitrary alternative count data models, such as hurdle or zero-inflated models where the conditional expectation function is more complex.

IV Estimation With Multiplicative Errors

We consider the model

$$\mathrm{E}(y|x,u) = \exp(\alpha + x\beta)u \tag{5.54}$$

where $\mathrm{E}(u|x)$ is a function of x. We explicitly keep track of the constant for reasons that will become apparent shortly. Also, we use, for simplicity, a single endogenous regressor x. With endogeneity, $\mathrm{E}(y|x)$ is not proportional to $\exp(\alpha + x'\beta)$, and the Poisson PML for β is inconsistent.

Assume that an instrument z is available such that $\mathrm{E}(u|z) = \delta$ where δ is some constant. Then

$$\mathrm{E}[y\exp(-\alpha - x\beta)|z] = \delta$$

or, equivalently,

$$\mathrm{E}[y\exp(-\tilde{\alpha} - x\beta) - 1|z] = 0 \tag{5.55}$$

where $\tilde{\alpha} = \alpha + \log\delta$. The moment restrictions depend on $\tilde{\alpha}$ and on β. δ is of course unknown, so that we cannot recover the original intercept unless we simply assume that $\delta = 1$. (5.55) implies any number of unconditional moment conditions

$$\mathrm{E}[(y\exp(-\tilde{\alpha} - x\beta) - 1)g(z)] = 0$$

for arbitrary functions $g(z)$.

Such derived moment conditions can be used to estimate β consistently (Mullahy, 1997a). In particular, for $g(z) = z$, and provided that a sufficient number of instruments is available, estimation can be based on the empirical covariance between $u - 1$ and z, i.e.,

$$\sum_{i=1}^{n}(y_i\exp(-\tilde{\alpha} - x_i\beta) - 1)z_i = 0 \tag{5.56}$$

If more instruments than endogenous variables are present, GMM estimation is asymptotically efficient, and tests for over-identification can be implemented. (See, for instance, Davidson and MacKinnon, 1993, Windmeijer and Santos Silva, 1997).

IV Estimation With Additive Errors

Let

$$\mathrm{E}(y|x,\eta) = \exp(\alpha + x\beta) + \eta$$

x is endogenous if $\mathrm{E}(\eta|x)$ is a function of x. Thus

$$\mathrm{E}[(y - \exp(\alpha + x\beta)|x] \neq 0$$

and Poisson regression is inconsistent. However, suppose that instruments z are available such that $\mathrm{E}(\eta|z) = 0$. Because of the additive separability, the normalization of the conditional expectation to zero is inconsequential. We obtain

$$\mathrm{E}[y - \exp(\alpha + x\beta)|z] = 0 \tag{5.57}$$

from where we can derive any number of moment restrictions. In particular, it implies zero correlation between η and z. Grogger (1990b) discusses this approach with special reference to count data modeling. He points out that a comparison of the non-linear instrumental variable estimator with the Poisson maximum likelihood estimator provides a simple test for exogeneity of the regressors.

As shown by Windmeijer and Santos Silva (1997), the same instrument cannot, in general, be valid in the multiplicative and additive case alike. Recall that

$$\begin{aligned}\mathrm{E}(y|x,u) &= \exp(\alpha + x\beta)u \\ &= \exp(\alpha + x\beta) + \exp(\alpha + x\beta)(u-1)\end{aligned}$$

Assume that $\mathrm{E}(u|z) = 1$ (or some arbitraty constant) so that z is a valid instrument in the multiplicative model. Then $\mathrm{E}[y - \exp(\alpha + x\beta)|z] \neq 0$, in general, since $\mathrm{E}(\exp(\alpha+x\beta)(u-1)|z) = \mathrm{E}(\exp(\alpha+x\beta)u|z) - \mathrm{E}(\exp(\alpha+x\beta)|z)$. But the first term does not simplify in general due to the correlation between z and x.

There are exceptions. Windmeijer (2008) shows that if endogeneity is due to classical measurement error in an explanatory variable, then both additive and multiplicative moment conditions are valid. This is also the case if x has a linear first stage,

$$x = z'\gamma + \varepsilon$$

where z is independent of ε and u. With

$$\mathrm{E}(y|x,u) = \delta \exp(\alpha + x\beta) + \exp(\alpha + x\beta)(u - \delta)$$

we obtain

$$\mathrm{E}[y - \delta \exp(\alpha + x\beta)|z] = \exp(\alpha + \beta(z'\gamma))[\mathrm{E}(\exp(\beta\varepsilon)u - \delta\mathrm{E}(\exp(\beta\varepsilon)]$$

which is zero for $\delta = \mathrm{E}(\exp(\beta\varepsilon)u)/\mathrm{E}(\exp(\beta\varepsilon))$. We obtain a valid linear IV estimator based on the moment condition

$$\mathrm{E}[y - \exp(\tilde{\alpha} + x\beta)|z] = 0$$

where $\tilde{\alpha} = \log \mathrm{E}(\exp(\beta\varepsilon)u - \log \mathrm{E}(\exp(\beta\varepsilon)$.

5.3.6 Estimation in Stages

To understand, when and why two-step estimation, or estimation in stages, may work in count data models, we start from the conditional expectation function with multiplicative error. For simplicity, we consider again a model with a single endogenous regressor

$$\mathrm{E}(y|x,u) = \exp(\alpha + x\beta)u$$

Assume that x has a first stage equation

$$x = z'\gamma + \varepsilon \tag{5.58}$$

where the instruments z are *fully independent* of ε and u. This assumption is obviously stronger than mean independence, or mere lack of correlation. This assumption rules out, for instance, that x is a binary endogenous regressor (see below) or a non-negative regressor.

Endogeneity arises because $\mathrm{Cov}(u,\varepsilon) \neq 0$. Under this set of assumptions,

$$\begin{aligned}\mathrm{E}(y|z) &= \mathrm{E}[\exp(\alpha + x\beta)u|z] \\ &= \mathrm{E}[\exp(\alpha + \beta(z'\gamma + \varepsilon))u|z] \\ &= \exp(\alpha + \beta z'\gamma)\mathrm{E}[\exp(\beta\varepsilon)u|z]\end{aligned}$$

It follows from independence that $\mathrm{E}(\exp(\beta\varepsilon)u|z)$ is a constant not depending on z and therefore

$$\mathrm{E}(y|z) = \exp(\tilde{\alpha} + \beta z'\gamma) \tag{5.59}$$

where $\tilde{\alpha} = \alpha + \log \mathrm{E}(\exp(\beta\varepsilon)u)$. $\tilde{\alpha}$ and β can be estimated consistently based on the moment condition

$$\mathrm{E}(y - \exp(\tilde{\alpha} + \beta z'\hat{\gamma})|z) = 0 \tag{5.60}$$

where the unknown γ has been replaced by a consistent first stage estimator. Thus, in practice, one can perform a Poisson estimation, regressing y on $\hat{x} = z'\hat{\gamma}$ where $\hat{\gamma}$ is obtained from a first-stage estimation. Estimation in stages does not identify the true constant of the model, α. As a consequence, it does not identify any quantity of interest that requires knowledge of this constant. One example for such a non-identified quantity is the partial mean effect for a given u. However, all relative mean effects are identified, as is the slope parameter β.

This approach requires an exclusion restriction (z may not have a direct effect on y). Also, note that in this non-linear set-up, two-stage estimation and non-linear instrumental variables estimation do not amount to the same thing. The standard errors of the second stage estimates $\hat{\tilde{\alpha}}$ and $\hat{\beta}$ need to be adjusted to account for the sampling variation introduced from the estimation of γ, which can be done using standard formulas for two-step estimation (Murphy and Topel, 1985). Alternatively, one can bootstrap the standard errors.

Wooldridge (2002) discusses a slightly different form of estimation in stages for count data models with endogenous regressors. Point of departure is a

conditional mean function with multiplicative unobserved heterogeneity as before,

$$\mathrm{E}(y|x,v) = \exp(\alpha + \beta x + v)$$

where $v = \log u$ and

$$x = z'\gamma + \varepsilon$$

as before. In addition, assume that

$$v = \rho\varepsilon + \xi \tag{5.61}$$

where ξ is independent of ε. Note that the linear conditional expectation assumption is made for v, not for u. This formulation corresponds to the notion that endogeneity derives from omitted variables, which means that v and x should be treated symmetrically. It follows then that

$$\mathrm{E}(y|x,\varepsilon) = \exp(\tilde{\alpha} + \beta x + \rho\varepsilon)$$

where $\tilde{\alpha} = \alpha + \log \mathrm{E}[\exp(\xi)]$. While ε is not observed it can be estimated as residual from a first stage regression of x on z. Estimates for the parameters of the exponential regression, $\hat{\tilde{\alpha}}$, $\hat{\beta}$, and $\hat{\rho}$ can then be obtained from a Poisson quasi likelihood regression. The Wooldridge approach requires one additional assumption, (5.61). In return, it is possible to estimate a model for $\mathrm{E}(y|x,\varepsilon)$, rather than $\mathrm{E}(y|z)$, which has the advantage that ρ is estimated explicitly, and the hypothesis of endogeneity can be directly tested.

Note that both approaches require linearity and independence assumptions. This is an important difference to the linear case. Here, the linear reduced form model for the right hand side endogenous variable actually amounts to an additional assumption that is required (Mullahy, 1997a, Wooldridge, 1997b). The second stage Poisson estimators using first stage fitted values is inconsistent unless the model for the first stage conditional expectation function is correct (and ε is fully independent of z). On the other hand, conventional two-stage least squares estimators using a linear second stage model are consistent whether or not the first stage conditional expectation function is linear. In fact, two-stage least squares is equivalent to instrumental variable estimation then.

Thus, there is some pay-off to actually avoiding the exponential mean function. If one is willing to specify the Poisson regression model with linear mean function, two-stage estimation is consistent under considerably weaker assumptions on the reduced form equation (See, for instance, Mullahy and Portney, 1990, and also Sander, 1992). Angrist (2001) uses a related argument in the context of a linear outcome equation combined with a binary endogenous regressor to argue for using a linear probability model at the first stage.

Binary Endogenous Regressor

Estimation in stages does not work always, and in particular not, when x is a binary endogenous variable. Two-stage estimation of a count model with endogenous binary variable has also been referred to as "forbidden regression" (Wooldridge, 2002). The reason for its failure to deliver a consistent estimator is that for a binary endogenous variable d, there does not exist a reduced form representation with fully independent error. In a binary model, the error term is heteroskedastic by construction, and thus not independent of z. But this is a key requirement for the whole approach. For example, if we replace a single binary endogenous regressor by its predicted probability, we can obtain an expression such as

$$\mathrm{E}(y|z) = \exp(\beta_0 + \beta_1 F(z'\gamma))\mathrm{E}[\exp(\beta_1 w)|z] \tag{5.62}$$

where $w = d - F(z'\gamma)$. Now, $\mathrm{E}[\exp(\beta_1 w)|z]$ is a function of z. The reason is that the higher order moments of w depend on z. For example, $\mathrm{E}(w^2|z) = F(z'\gamma)[1 - F(z'\gamma)]$. But $\mathrm{E}(\exp(\beta_1 w))$, because of the convexity of the exponential transformation, is an increasing function of the variance, and therefore depends on both parameters and regressors in this case.

5.4 Switching Regression

So far, we have considered non-linear instrumental variable estimation and estimation in stages. A further estimation method for models with endogenous regressors becomes available if one is willing to specify the full joint distribution of y_1 and y_2. This goes back to the initial endogeneity definition by Engle, Hendry and Richard (1983) (see Chap. 5.3.2) where it was pointed out that structural parameters can always be estimated from a fully specified joint model. We will discuss such *full information maximum likelihood estimation* based on a parametric bivariate model within the context of the switching regression model, or its close cousin – a count data model with endogenous binary regressor – mainly, because this has been the main application in the literature (Terza, 1998)

The switching regression model can be summarized by the following three equations:

$$y_0 = \exp(x'\beta_0 + v_0) + u_0$$

$$y_1 = \exp(x'\beta_1 + v_1) + u_1$$

$$d = \mathbb{1}(z'\gamma + \varepsilon > 0)$$

and the observation rule that y_1 is observed when $d = 1$ and y_0 is observed when $d = 0$. Endogeneity enters the model via the unobserved heterogeneity terms v_0 and v_1, respectively. Specifically, assume that v_0 and ε are correlated,

as are v_1 and ε. On the other hand, the additional equation errors u_0 and u_1 are independent of v_0, v_1, ε and x. If we restrict $u_0 = u_1$, $v_0 = v_1$, and $\beta_0 = \beta_1$ except for the constant, we obtain the model with endogenous binary regressor

$$y = \exp(x'\beta + \alpha d + v) + u \tag{5.63}$$

where d is determined as before.

5.4.1 Full Information Maximum Likelihood Estimation

We focus here on the full information maximum likelihood approach for the model with binary endogenous variable (Terza, 1998). The approach for the full switching regression model would be very similar, as it would for a model with any arbitrarily scaled endogenous regressor. Note that the Terza model is also closely related to other non-linear models with binary endogenous regressors, such as the bivariate probit model as described, for instance, by Evans and Schwab (1995).

To obtain the full joint parametric model, we assume that $f(y|x, d, v)$ in (5.63) is a standard count data model (for example a Poisson distribution). Moreover, v and ε are independent of x and z, but correlated with each other. This correlation is the source of endogeneity in this type of model. The typical interpretation is the existence of unobserved explanatory variables, that affect both the rate at which events occur and the probability of being observed in the state $d = 1$.

The full information maximum likelihood approach requires a parametric assumption on the joint distribution of v and ε. The observed data identify $f(y, d|x, z)$. We can write

$$\begin{aligned} f(y, d|x, z) &= f(y|d, x, z) f(d|x, z) \\ &= \int_{-\infty}^{\infty} f(y|d, x, z, v) f(d|x, z, v) g(v) dv \end{aligned}$$

The term under the integral is a product of two conditional and one marginal distribution. The first conditional distribution is fully specified, for instance a Poisson model. The second conditional distribution is a binary model for d that is now conditional on v, thus determined by the conditional distribution of $\varepsilon|v$. The model is fully specified once we know $f(\varepsilon|v)$ and $g(v)$ (and thus the joint distribution of the two error terms).

The common assumption is, of course, that v and ε have a bivariate normal distribution with mean zero, variance σ^2 and 1, respectively, and correlation parameter ρ. This case is relatively easy to handle, and it captures the dependence between the two equations in a single parameter, ρ that varies between -1 and +1.

Under these assumptions,

$$f(d|z, v) = P(d = 1|z, v)^d [1 - P(d = 1|z, v)]^{1-d}$$

where

$$P(d=1|z,v) = P(\varepsilon > -z'\gamma|v) \\ = \Phi^*(v,z) \tag{5.64}$$

where

$$\Phi^*(v,z) = \Phi\left(\frac{z'\gamma + \rho v/\sigma}{\sqrt{1-\rho^2}}\right)$$

and Φ denotes the cumulative density function of the standard normal distribution (see 5.32). Finally, $g(v)$ is a normal distribution with mean 0 and variance σ^2.

Collecting all terms, the full joint probability function of the count data model with endogenous binary regressor can be written as

$$f(y,d|x,z) = \int_{-\infty}^{\infty} f(y|d,x,v)\Phi^*(v,z)^d[1-\Phi^*(v,z)]^{1-d}g(v)\mathrm{d}v \tag{5.65}$$

While the integral has no closed-form solution, numerical approximation using quadrature or other simulation methods provides no major difficulties. The parameters can be estimated by maximizing the log-likelihood function of the sample

$$\ell(\beta,\gamma,\sigma^2,\rho) = \sum_{i=1}^{n} \log f(y_i,d_i|x_i,z_i;\beta,\gamma,\sigma^2,\rho)$$

with respect to β, γ, σ^2 and ρ. This estimator has all the useful properties of maximum likelihood estimators provided the model is correctly specified.

A Bayesian analysis of this model is provided by Kozumi (2002). Deb and Trivedi (2006) extend the model to allow for an endogenous multinomial variable, while keeping a normality assumption (among random errors in a mixed multinomial logit model and the unobserved heterogeneity in the count process).

What are the consequences of ignoring endogeneity of d? From (5.63), we can deduce that

$$\frac{\mathrm{E}(y|x,d=1)}{\mathrm{E}(y|x,d=0)} = \frac{\mathrm{E}_v\mathrm{E}(y|x,v,d=1)}{\mathrm{E}_v\mathrm{E}(y|x,v,d=0)} \\ = \frac{\exp(x'\beta+\alpha)\mathrm{E}(\exp(v)|d=1)}{\exp(x'\beta)\mathrm{E}(\exp(v)|d=0)}$$

In order to evaluate the expectations, we need to make reference to the aforementioned results on truncation in the log-normal distribution (see equation (5.28)). In particular,

$$\mathrm{E}(\exp(v)|d=1) = \exp(\sigma^2/2)\frac{\Phi(z'\gamma+\rho\sigma)}{\Phi(z'\gamma)}$$

and

$$\mathrm{E}(\exp(v)|d=0) = \exp(\sigma^2/2)\frac{\Phi(-z'\gamma-\rho\sigma)}{\Phi(-z'\gamma)}$$

Therefore, under the assumptions of this model

$$\frac{\mathrm{E}(y|x,d=1)}{\mathrm{E}(y|x,d=0)} = \exp(\alpha)\frac{\Phi(z'\gamma+\rho\sigma)}{\Phi(z'\gamma)}\frac{\Phi(-z'\gamma)}{\Phi(-z'\gamma-\rho\sigma)} \tag{5.66}$$

If $\rho > 0$, it is easily verified that the factor following $\exp(\alpha)$ is always greater than 1. In other words, the overall relative difference between the two expected counts exceeds then $\exp(\alpha) - 1$, the causal difference that would be observed for two randomly selected, otherwise identical individuals, for one of whom $d = 1$ and for the other $d = 0$. Ignoring the endogeneity of d therefore leads to an upward bias in the estimated effect for $\rho > 0$. For $\rho < 0$, there is a downward bias.

5.4.2 Moment-Based Estimation

Terza (1998) also discusses estimation of the model under weaker assumptions.

- $f(y|d,v)$ is not specified (for instance no assumption of a Poisson or negative binomial distribution). Only the conditional expectation function is specified:

  $$\mathrm{E}(y|d,v) = \exp(x'\beta+\alpha d+v)$$

- ε and v are bivariate normal distributed as before.

We know that in this case

$$\mathrm{E}(y|x,z,d=1) = \exp(x'\beta+\alpha+\sigma^2/2)\frac{\Phi(z'\gamma+\rho\sigma)}{\Phi(z'\gamma)}$$

and

$$\mathrm{E}(y|x,z,d=0) = \exp(x'\beta+\sigma^2/2)\frac{\Phi(-z'\gamma-\rho\sigma)}{\Phi(-z'\gamma)}$$

Terza (1998) suggests a two stage estimation method similar to that proposed by Heckman (1979) for the linear model.

a) Estimate a probit model and obtain consistent estimates $\hat{\gamma}$ for γ.
a) Estimate a regression model with multiplicative correction factor (5.28) by non-linear least squares. NLS is required, since the conditional expectation is non-linear in ρ, σ, and β.

It is unlikely that this estimation method will supersede the relatively straightforward full information maximum likelihood estimation in practice. The gain of a certain robustness, because one does not need to make a distributional assumption for the conditional distribution of the count, is at the same time a loss, since this is not a generic count data model any longer, where inferences on probabilities of single outcomes are possible. Moreover, the Poisson distribution has substantial robustness properties, so that one should feel in practice quite comfortable with this assumption, whereas the choice of the bivariate mixing distribution is potentially less innocuous. But it is exactly the bivariate normal assumption that Terza's moment estimator relies heavily on as well.

5.4.3 Non-Normality

Clearly, one might be concerned with the heavy reliance of this approach on particular distributional form assumptions. While the normal distribution certainly has a few arguments speaking in its favor, alternative approaches have been pursued in the literature, if only in order to provide tools for assessing the robustness of the results with respect to these assumptions. The issues here are essentially the same as those discussed in connection with sample selection models in Chap. 5.2.2. In fact, the formal structure of sample selection models, switching regression models, and models with endogenous binary regressor is very similar. In a sample selection model, the outcome is only observed in one state, whereas in a switching regression model, there is an outcome in each state, but the counterfactual outcome (for example y_1 if $d = 0$) is unobserved as well.

Summarizing the discussion of Chap. 5.2.2, previous proposals to distributions other than the multivariate normal have been based on copula functions (Ophem, 2000; see also Chap. 7.1.7), discrete factor approximation (Mroz, 1999), and using series expansions to obtain flexible form models for the conditional probability function of the counts, thereby implicitly acconting for the non-standard conditional expectation functions resulting from endogeneity (Romeu and Hernandez, 2005).

5.5 Mixed Discrete-Continuous Models

Models with bivariate normal unobserved heterogeneity structure can be easily extended to deal with a continuous endogenous variable. In this case, the joint density function can be written as

$$f(y_1, y_2) = \int f(y_1|y_2, v) f(y_2|v) g(v) d\varepsilon$$

If $f(y_1|y_2, v)$ is a count data distribution function, y_2 has reduced form

$$y_2 = z'\gamma + \varepsilon$$

and v and ε are bivariate normal, the conditional distribution of $y_2|v$ is a normal distribution with mean $z'\gamma + \rho\sigma_\varepsilon/\sigma_v v$ and variance $\sigma_\varepsilon^2 - \rho^2/\sigma_v^2$. Thus, it is relatively simple to establish the likelihood function. Of course, this approach is plausible only if y_2 is a continuous variable that can be (at least approximately) normal distributed.

A related class of models arises if two dependent variables, one of them possibly a count, are only connected through correlated errors. Thus, in the two equation framework of Chap. 5.3.4, both α and δ are zero, but the errors v and ε are correlated. Accounting for such correlation is then not an issue of endogeneity and consistent estimation but rather one of efficient estimation. In the traditional linear model terminology, one would refer to this set-up as one of *seemingly unrelated regressions*.

An example in the literature is Prieger (2002) who considers two outcome variables, the number of innovations in the telecommunication markets (a discrete variable modeled by a count data model) and the time until adoption in the market, or regulatory delay (a continuous variable modeled by a duration model, here a Weibull distribution). In such an application, the correlation ρ may have a substantive interpretation, providing evidence on possible congestion effects, as an increased number of innovations may for instance increase the regulatory delay due to congestion. Other interesting uses of such a model are conceivable, such as the joint modeling of the individual number of unemployment spells and their durations.

6

Zeros in Count Data Models

6.1 Introduction

There are two main reasons why zeros are of particular interest in count data models. First, empirically, their fraction is often too high to be compatible with a standard underlying count data model (we also speak of excess zeros then). Second, theoretically, zeros often reflect corner solution outcomes in economic choice models. In such cases, the process generating zeros might depend on other driving forces than the process for strictly positive outcomes, making it informative and relevant to distinguish between elasticities at the intensive and extensive margins. This issue is reminiscent of the debate in the limited dependent variable literature on the appropriateness of the Tobit model as opposed to so-called two-part models (Cragg, 1971, Duan et al., 1993).

The following examples illustrate the sense in which zeros are potentially different.

- Consider the study of individual fertility, measured as the number of births by a woman. The outcome "no births" can be due either to infertility or to choice.
- In the study of recreational demand, the number of trips to a ski field during the last quarter can be zero either because a person is not a skier, or because a skier did not go skiing during the last quarter. In this example, we are confronted with the situation that the time frame (a quarter) may be too short to observe low frequency events, despite the fact that latent demand is not zero.
- In modeling the number of job changes using worker survey data, some zeros will arise simply due to underreporting. Thus, in count data, zeros naturally result from measurement error in the dependent variable.
- In modeling health care utilization, it has been hypothesized that zero visits to a doctor reflect the state of health, whereas the positive number of visits is partially supply (i.e., physician) induced.

6.2 Zeros in the Poisson Model

The presence of zeros, even a very high fraction of zeros, is a well-defined outcome under the Poisson distribution and thus does not *per-se* rule out that model, or any other of the standard count data models discussed so far. For the Poisson distribution

$$f_{psn}(0) = e^{-\lambda} \tag{6.1}$$

(which is equal to the survivor probability of an exponentially distributed waiting time at $T = 1$). Thus, for the Poisson distribution, the probability of a zero and the overall mean are in a one-to-one inverse proportional relationship. To give an example of the magnitudes involved, a zero probability $f(0) = 0.5$ implies $\lambda = -\log(0.5) = 0.69$, whereas it follows from $f(0) = 0.25$ that $\lambda = 1.38$. A large fraction of zeros, and therefore a small mean, is thus fully compatible with the standard Poisson model. The primary consequence is that the data are relatively uninformative for estimating parameters, as the standard errors are a decreasing function of the sample mean. In the Poisson regression model, we have

$$\widehat{\mathrm{Var}}(\hat{\beta}) = \left[\sum_{i=1}^{n} \exp(x_i'\hat{\beta}) x_i x_i'\right]^{-1}$$

(see equation (3.34)) which becomes larger as $\exp(x_i'\hat{\beta})$ becomes smaller.

6.2.1 Excess Zeros and Overdispersion

In many cases, a large proportion of zeros will not only affect the precision of inference, but rather speak directly against the Poisson regression model. This is always the case when there are "too many" zeros ($f(0) > e^{-\lambda}$) or, less common in practice, "too few" zeros ($f(0) < e^{-\lambda}$). It turns out that excess zeros can be accommodated by the negative binomial model or, in fact, by any Poisson mixture model, as the following results show. We start with the negative binomial model in Negbin II specification. In this case

$$f_{nb}(0) = \left(\frac{\alpha}{\lambda + \alpha}\right)^{\alpha} = \frac{1}{\left(1 + \frac{\lambda}{\alpha}\right)^{\alpha}} \tag{6.2}$$

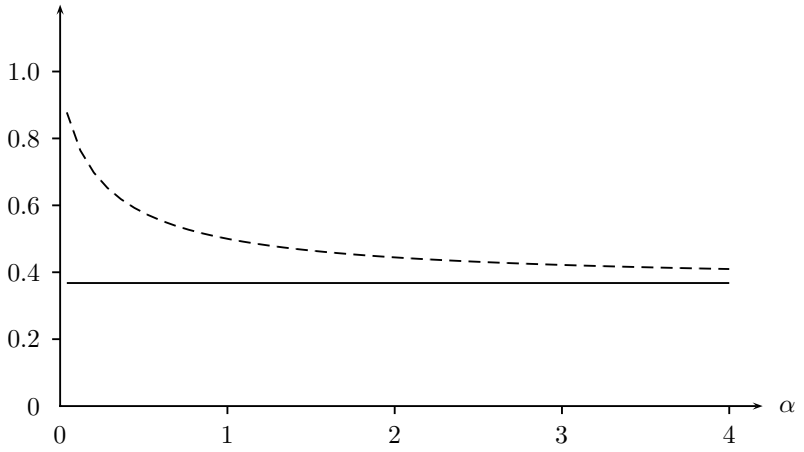
We see that $\lim_{\alpha \to \infty} f_{nb}(0) = f_{psn}(0)$ as required. For finite α, it holds that

$$\left(1 + \frac{\lambda}{\alpha}\right)^{\alpha} < e^{\lambda}$$

and therefore $f_{nb}(0) > f_{psn}(0)$. This relationship is also seen in Fig. 6.1.

Thus, a comparison between the Poisson distribution and the negative binomial distribution suggests a link between, unobserved heterogeneity, overdispersion and excess zeros. This result generalizes in a number of ways. First, unobserved heterogeneity must not necessarily be of the gamma type, as

Fig. 6.1. Probability of a Zero as a Function of α, for $\lambda = 1$, in Poisson (Solid Line) and Negative Binomial Distribution (Dashed Line)



is the case when generalizing the Poisson distribution to the negative binomial distribution. It is easy to see that any multiplicative unobserved heterogeneity in the Poisson model will generate excess zeros. If y is Poisson distributed with parameter λv, and v has density function $g(v)$, with $\mathrm{E}(v) = 1$, then

$$f(0|\lambda) = \int_0^{-\infty} e^{-\lambda v} g(v) dv = \mathrm{E}_v \left(e^{-\lambda v}\right)$$

Since the exponential function is a convex function we have, by Jensen's inequality, that

$$\mathrm{E}_v \left(e^{-\lambda v}\right) \geq e^{-\lambda \mathrm{E}(v)} = e^{-\lambda}$$

where equality holds only when $\sigma_v^2 = 0$.

Second, one can consider distributions other than the Poisson and ask whether unobserved heterogeneity leads to excess zeros there as well. At least for any exponential family distribution, this is the case indeed, as the *Two-Crossings Theorem* due to Shaked (1980) establishes.

6.2.2 Two-Crossings Theorem

Suppose a base probability model $f(y|x, v)$ for y discrete or continuous is from the exponential family, and let $\mathrm{E}(v) = 1$ and $\mathrm{Var}(v) = \sigma_v^2 > 0$. Then the mixed distribution $f(y|x) = \mathrm{E}_v[f(y|x, v)]$ has heavier tails than the base distribution in the sense that the sign pattern of $f(y|x) - f(y|x, v = 1)$ is $\{+, -, +\}$ as y increases on its support.

The two-crossings theorem formalizes the intuition that mixing spreads out a distribution away from its center and towards its tails. While parametric models with overdispersion therefore offer one approach to account for extra zeros (relative to the Poisson benchmark), this may not always be the best one to follow. The main limitation remains, namely that zeros are generated from the same underlying process as positives. The rest of the chapter mostly deals with so-called *multi-index models*, where the zero-generating process is not subject to such a constraint.

6.2.3 Effects at the Extensive Margin

Count data models lend themselves naturally to address the question how a regressor induced change in λ affects the probability of a zero. In the Poisson model

$$\frac{df(0|\lambda)}{d\lambda} = -e^{-\lambda}$$

whereas in the negative binomial model

$$\frac{\partial f(0|\lambda,\alpha)}{\partial \lambda} = -\left(\frac{\alpha}{\lambda+\alpha}\right)^{\alpha+1}$$

Thus, the effect at the extensive margin is negative in either case. To obtain the effects of a change in a regressor, for $\lambda = \exp(x'\beta)$, the above expressions have to be multiplied by

$$\frac{\partial \lambda}{\partial x_j} = \lambda \beta_j$$

An increase in x_j reduces the probability of a zero outcome if $\beta_j > 0$, and increases it otherwise. Note that in the case of the Poisson model with unobserved heterogeneity, mixing does not have an unambiguous effect on the size of the marginal effect. We can write

$$\frac{df(0|\lambda)}{d\lambda} = \int_0^\infty \frac{df(0|\lambda,v)}{d\lambda} g(v)dv = \int_0^\infty -v\exp(-\lambda v)g(v)dv$$

which is the first derivative of the Laplace transform of v, evaluated at $s = \lambda$. While this expression must be negative, it is unclear whether it is greater or less than $-\exp(-\lambda)$ for arbitrary $g(v)$.

The important thing to note is another one, though. The marginal effects for the zero outcome are a function of λ, the mean of the count distribution. So upward shifts in the mean must reduce the probability of a zero, in the Poisson case following an exponential decay pattern. This is very restrictive. Even if the sign may be less disputed, the true magnitude could differ from the one imposed by functional form.

Consider, as an example, health care utilization, measured in terms of number of doctor visits. The factors that explain whether a doctor is contacted a first time may differ from the factors that determine the number

of visits (re-appointments, referrals) following a first visit. Similarly, for the fertility example mentioned earlier, the factors behind infertility are presumably different than those determining choice. And even if the factors were the same, we may want to consider models, in which the quantitative response at the intensive and extensive margins can be estimated freely, i.e., determined from the data, rather than being *a-priori* determined by functional form.

6.2.4 Multi-Index Models

In order to obtain such additional flexibility, we essentially have to move beyond the class of single-index models and consider more general multi-index models instead. In the remainder of this chapter, three classes of multi-index models are distinguished and discussed in some detail. These are

- Hurdle Models
- Zero Inflated Models
- Compound Poisson Models

All three classes of models address, in their specific ways, the possibility that the distributional response to a variation in an exogenous variable may be different in different parts of the outcome distribution, and in particular at the extensive (the zeros) and intensive (the positives) margins, relative to what a standard count data model, such as the Poisson or the negative binomial model, would imply. At the end of this chapter, we add a discussion of a fourth such generalization, namely quantile regression for counts. The quantile approach is similar to the three probability based ones, listed above, as it also has a multi-index structure and therefore allows for different response elasticities in different parts of the outcome distribution. At the same time, it is fundamentally different, since it focuses on quantiles rather than on probabilities. The two views are complementary. While quantiles and probabilities are not in a one-to-one relation, both approaches should lead to qualitatively similar results. For example, from the quantile perspective, one could define the extensive margin in terms of a low quantile, such as the first decile, rather than in terms of the zero outcome, as it would be appropriate in a probability model.

6.2.5 A General Decomposition Result

The probability models discussed in this chapter all have conditional mean functions that depart from the simple log-linear mean of the standard Poisson regression model. Thus, parameters cannot be interpreted directly as proportional mean effects. The overall mean is usually not the main object of interest in this class of models. Rather, any mean effect can be decomposed into an effects at the intensive margin, and an effect at the extensive margin, i.e. for $P(y = 0|x)$ and for $\mathrm{E}(y|y > 0, x)$, respectively. Winkelmann (2004b) reports

results of such a decomposition in the context of an evaluation of a health care reform in Germany, and its effect on the individual number of visits to a physician. In that study, it was hypoythesized that the reform might have had a disproportionately large effect at the left end of the outcome distribution, and for zero visits in particular. Tests based on multi-index models did confirm such an asymmetric response.

Formally, we can always write

$$\mathrm{E}(y|x) = P(y > 0|x)\mathrm{E}(y|y > 0, x)$$

where $P(y > 0|x) = 1 - P(y = 0|x)$. Therefore

$$\frac{\partial \mathrm{E}(y|x)}{\partial x} = \frac{\partial P(y > 0, x)}{\partial x}\mathrm{E}(y|y > 0, x) + \frac{\partial \mathrm{E}(y|y > 0, x)}{\partial x}P(y > 0|x)$$

or

$$\frac{\partial \mathrm{E}(y|x)/\mathrm{E}(y|x)}{\partial x} = \frac{\partial P(y > 0, x)/P(y > 0, x)}{\partial x} + \frac{\partial \mathrm{E}(y|y > 0, x)/\mathrm{E}(y|y > 0, x)}{\partial x}$$

As stated, such a decomposition is possible for any count data model. In standard models, however, it remains purely tautological. Both effects $\partial P(y > 0|x)/\partial x$ and $\partial \mathrm{E}(y|y > 0, x)/\partial x$ are functions of the same single index and the same parameter β.

The multi-index models discussed in this chapter severe this strict link, since $P(y > 0|x)$ and $\mathrm{E}(y|y > 0, x)$ are in general functions of different parameters. Marginal effects in these models are different in different parts of the outcome distribution, *relative* to a standard single index count data model. For example, in a multi index model, a given change in a regressor x can decrease the probability of a zero, but leave the conditional expectation $\mathrm{E}(y|y > 0, x)$ unchanged. The standard Poisson model would rule out such a possibility *a-priori*, based on functional form assumptions.

6.3 Hurdle Count Data Models

Hurdle count data models were first discussed by Mullahy (1986) (see also Creel and Loomis, 1990). Hurdle models allow for a systematic difference in the statistical process governing individuals (observations) below the hurdle and individuals above the hurdle. In particular, a hurdle model combines a dichotomous model for the binary outcome of the count being below or above the hurdle (the selection variable), with a truncated model for outcomes above the hurdle. For this reason, hurdle models sometimes are also referred to as two-part models.

The most widely used hurdle count data model sets the hurdle at zero. Only this model is discussed in this chapter. From a statistical point of view, the hurdle at zero formulation can account for excess zeros, as defined in

Chapter 6.2.1. From an economic point of view, an intuitive appeal arises from its interpretation as a two-part decision process, a plausible feature of individual behavior in many situations.

The hurdle model is flexible and allows for both under- and overdispersion in a way to be discussed below. Applications include Pohlmeier and Ulrich (1995), Arulampalam and Booth (1997), and Booth, Arulampalam and Elias (1997) who estimate hurdle negative binomial models for the determinants of visits to a physician and the incidence of training, respectively. Gurmu and Trivedi (1996) apply a hurdle model to the annual number of recreational boating trips by a family. Wilson (1992) proposed a hurdle Poisson model with endogenously determined position of the hurdle.

For a general formulation, assume that $g_1(0)$ is the probability of a zero outcome, and that $g_2(k)$, $k = 1, 2, \ldots$ is a probability function for positive integers. The probability function of the hurdle-at-zero model is then given by:

$$
\begin{aligned}
f(y = 0) &= g_1(0) \\
f(y = k) &= (1 - g_1(0))g_2(k) \; k = 1, 2, \ldots
\end{aligned}
$$

Mullahy (1986) advocates an approach where both parts of the hurdle model are based on probability functions for nonnegative integers, call them f_1 and f_2, respectively. In terms of the general model above, let $g_1(0) = f_1(0)$ and $g_2(k) = f_2(k)/(1 - f_2(0))$. In the case of g_2, a normalization is required since f_2 has support over the nonnegative integers whereas the support of g_2 must be over the positive integers. Formally, this corresponds to truncation of f_2. However, there is no truncation of the population here. All that is needed is a distribution with positive support, and the second part of a hurdle model can use a displaced distribution, or any distribution with positive support as well.

Under the Mullahy (1986) assumptions, the probability distribution of the hurdle-at-zero model is given by

$$
\begin{aligned}
f(y = 0) &= f_1(0) \\
f(y = k) &= \frac{1 - f_1(0)}{1 - f_2(0)} f_2(k) \\
&= \Theta f_2(k), \qquad k = 1, 2, \ldots
\end{aligned}
\tag{6.3}
$$

where f_2 is referred to as *parent*-process. The numerator of Θ gives the probability of crossing the hurdle and the denominator is a normalization that accounts for the (purely technical) truncation of f_2. It follows that the hurdle model collapses to the parent model if $f_1 = f_2$ or, equivalently, $\Theta = 1$. The expected value of the hurdle model is given by

$$
\mathrm{E}_h(y) = \Theta \sum_{k=1}^{\infty} k f_2(k) = \Theta \mathrm{E}_2(y)
$$

It differs from the expected value of the parent model by a factor Θ. If the probability of crossing the hurdle is greater than the sum of the probabilities of positive outcomes in the parent model, Θ exceeds 1, thus increasing the expected value of the hurdle model relatively to the expected value of the parent model. Alternatively, if the probability of not crossing the hurdle is greater than the probability of a zero in the parent model – the usual case in an application with excess zeros – Θ is less than 1, thus decreasing the expected value of the hurdle model relatively to the expected value of the parent model. This model thus provides a new interpretation of excess zeros as being a feature of the mean function rather than a feature of the variance function. The mean function of the hurdle model introduces additional non-linearities relative to the standard model in order to account for the corner solution outcome, much as in other corner solution models, such as for instance the Tobit model.

In addition, the hurdle model leads to a modified variance to mean ratio. The variance is

$$\mathrm{Var}_h(y) = \sum_{k=1}^{\infty} k^2 f_2(k)\Theta - \left[\Theta \sum_{k=1}^{\infty} k f_2(k)\right]^2$$

and the variance-mean ratio can be written as

$$\frac{\mathrm{Var}_h(y)}{\mathrm{E}_h(y)} = \frac{\sum_{k=1}^{\infty} k^2 f_2(k) - \Theta[\sum_{k=1}^{\infty} k f_2(k)]^2}{\sum_{k=1}^{\infty} k f_2(k)} \tag{6.4}$$

For $\Theta = 1$, (6.4) reduces to the variance-mean ratio of the parent model. If f_2 is a Poisson distribution function, this is equidispersion with $\mathrm{Var}(y)/\mathrm{E}(y) = 1$. For f_2 Poisson and $\Theta \neq 1$, (6.3) defines a *hurdle Poisson* model. $0 < \Theta < 1$ yields overdispersion, $1 < \Theta < c$ underdispersion. Mullahy (1986) sets $c = \infty$, but this does not hold in general since there is an upper limit to keep the variance positive. E.g., for the Poisson case

$$\sum_{k=1}^{\infty} k^2 f_2(k) = \lambda_2(\lambda_2 + 1)$$

where λ_2 is the expected value of the (untruncated) parent distribution. Hence

$$\mathrm{Var}_h(y) = \Theta\lambda_2(\lambda_2 + 1) - \Theta^2\lambda_2^2$$

with roots

$$\Theta_1 = 0\,, \qquad \Theta_2 = \frac{\lambda_2 + 1}{\lambda_2}$$

Thus, for the hurdle Poisson model underdispersion is obtained for $1 < \Theta < (\lambda_2 + 1)/\lambda_2$. For $\lambda_2 \to \infty$, underdispersion becomes impossible. This reflects the fact that underdispersion occurs if zeros are less frequent than the parent distribution would predict. The higher the expected value of the Poisson distribution, the lower the predicted probability of zero outcome and the lower the scope for underdispersion.

Double Hurdle Model

Hurdle models can be generalized to include more than a single hurdle. For example, let f_1, f_2, and f_3 be arbitrary probability distribution functions for non-negative integers. A double-hurdle model, or a model with two hurdles, in this example a first one at zero and a second one at one, has probability function

$$f(0) = f_1(0)$$

$$f(1) = \frac{1 - f_1(0)}{1 - f_2(0)} f_2(1)$$

$$f(k) = \left[1 - f_1(0) - \frac{1 - f_1(0)}{1 - f_2(0)} f_2(1)\right] \frac{f_3(k)}{1 - f_3(0) - f_3(1)}, \; k = 2, 3, \ldots$$

This is a straightforward generalization of (6.3). No applications of such a model are known so far.

Likelihood Function

The generic likelihood function for the hurdle-at-zero model with independent sampling is given by

$$L = \prod_{i=1}^{n} f_1(0; \theta_1)^{d_i} [1 - f_1(0; \theta_1)]^{1-d_i} [f_2(y_i; \theta_2)/(1 - f_2(0; \theta_2))]^{1-d_i} \qquad (6.5)$$

or, in logarithmic form,

$$\ell = \sum_{i=1}^{n} d_i \log f_1(0; \theta_1) + (1 - d_i) \log[1 - f_1(0; \theta_1)] + (1 - d_i) \log[f_2(y_i; \theta_2)/(1 - f_2(0; \theta_2))]$$

where $d_i = 1 - \min\{y_i, 1\}$. The first two terms on the right-hand side refer to the likelihood of the hurdle step, while the third term is the likelihood for positive counts. The log-likelihood of this parameterization is therefore separable, and maximization can be simplified by first maximizing a binary model log-likelihood using all observations, and then separately maximizing the log-likelihood for a truncated variable using the subset of observations for which the counts are positive.

6.3.1 Hurdle Poisson Model

Clearly, the hurdle model can be specified in a variety of ways by choosing different probability distributions f_1 and f_2 and specific parameterizations, like for instance Poisson, geometric, or negative binomial. Mullahy (1986) proposes the use of two Poisson distributions with $\lambda_1 = \exp(x'\beta_1)$ and $\lambda_2 =$

$\exp(x'\beta_2)$. This approach is convenient since the standard Poisson model can be tested via the parametric restriction $H_0 : \beta_1 = \beta_2$ using Wald or likelihood ratio test statistics.

The hurdle Poisson model has conditional expectation

$$\mathrm{E}(y|x) = \frac{(1 - e^{-\lambda_1})}{(1 - e^{-\lambda_2})}\lambda_2 \tag{6.6}$$

and variance function

$$\mathrm{Var}(y|x) = \mathrm{E}(y|x) + \frac{1 - \Theta}{\Theta}[\mathrm{E}(y|x)]^2 \tag{6.7}$$

where $\Theta = (1 - e^{-\lambda_1})/(1 - e^{-\lambda_2})$. It closely resembles the variance function (3.71) of the negative binomial model for $k = 1$, with the difference that the coefficient $\sigma^2 = (1 - \Theta)/\Theta$ now varies between individuals. The likelihood function of the hurdle Poisson model is given by

$$\begin{aligned} L(\beta_1, \beta_2) = \prod_{i=1}^{n} & \exp(-\exp(x_i'\beta_1))^{d_i}[1 - \exp(-\exp(x_i'\beta_1))]^{1-d_i} \\ & \times \left[\frac{\exp(-\exp(x_i'\beta_2))\exp(yx_i'\beta_2))}{y_i![1 - \exp(-\exp(x_i'\beta_2))]}\right]^{1-d_i} \end{aligned}$$

where $d_i = 1 - \min\{y_i, 1\}$.

6.3.2 Marginal Effects

Consider the marginal mean effect first. The conditional expectation function of the hurdle model was given in (6.6). Taking first derivatives, we obtain

$$\frac{\partial \mathrm{E}(y|x)}{\partial x} = \frac{\lambda_1 e^{-\lambda_1}}{1 - e^{-\lambda_2}}\lambda_2\beta_1 - \frac{1 - e^{-\lambda_1}}{(1 - e^{-\lambda_2})^2}\lambda_2^2 e^{-\lambda_2}\beta_2 + \frac{1 - e^{-\lambda_1}}{1 - e^{-\lambda_2}}\lambda_2\beta_2 \tag{6.8}$$

This partial effect is thus considerably more general (and complicated) than the partial effects of the single-index Poisson model, unless $\beta_1 = \beta_2$, in which case they simplify to the standard Poisson effects, $\partial \mathrm{E}(y|x)/\partial x = \lambda\beta$.

Similarly, we can compute the marginal probability effects for the hurdle Poisson model, at the extensive and at the intensive margins. We obtain

$$\begin{aligned} \frac{\partial f(0; x)}{\partial x} &= -\lambda_1 e^{-\lambda_1}\beta_1 \\ \frac{\partial f(k; x)}{\partial x} &= f(k; x)\left[\frac{\lambda_1 e^{-\lambda_1}}{1 - e^{-\lambda_1}}\beta_1 - \frac{\lambda_2 e^{-\lambda_2}}{1 - e^{-\lambda_2}}\beta_2 + (k - \lambda_2)\beta_2\right] \end{aligned} \tag{6.9}$$

where $k = 1, 2, \ldots$ and $f(k; x)$ is the probability function of the hurdle model. The single crossing restriction of the Poisson model does not apply here. It is softened by the introduction of a second parameter vector. In fact, it is clear from (6.9) that the marginal probability effects of the hurdle Poisson model can switch signs twice. So it is still not perfectly flexible, although more so than the simple Poisson model.

6.3.3 Hurdle Negative Binomial Model

By far the most popular hurdle model in practice is the hurdle-at-zero negative binomial model (Pohlmeier and Ulrich, 1995). In this case, $f_1 \sim NB(\beta_1, \alpha_1)$ and $f_2 \sim NB(\beta_2, \alpha_2)$. Estimation of the model can be based on the general likelihood factorization described above. This specification can give rise to an identification problem, as noted by Pohlmeier and Ulrich (1995).

Consider estimation of the hurdle part of the model, i.e., the parameters β_1 and α_1. This estimation is based on the dichotomous model $f_1(0; \beta_1, \alpha_1)$ versus $1 - f_1(0; \beta_1, \alpha_1)$. From (4.10), we see that for the generic negative binomial model

$$P(y = 0) = \left(\frac{\alpha_1}{\alpha_1 + \lambda}\right)^{\alpha_1}$$

The Negbin II model is obtained directly by letting $\lambda = \exp(x'\beta_1)$. For the Negbin I model,

$$P(y = 0) = \left(\frac{1}{1 + \alpha_1}\right)^{\lambda/\alpha_1}$$

In the Negbin I model, we can thus write

$$\begin{aligned} \log P(y = 0) &= \frac{\lambda}{\alpha_1} \log(1 + \alpha_1) \\ &= \exp(x'\beta_1 + \log \theta) \end{aligned}$$

where $\theta = \log(1 + \alpha_1)/\alpha_1 = f(\alpha_1)$. Hence, α_1 is not identified, as long as the regression part of the model contains a constant. In the Negbin II model, by contrast, two overdispersion parameters can be estimated, since

$$\log P(y = 0) = -\alpha_1 \log(1 + \alpha_1^{-1} \exp(x'\beta_1))$$

which ensures identification of $\sigma_1^2 = 1/\alpha_1$ based on functional form. In practice, however, this may be asking too much from the data, identification may be weak, and convergence problems can arise. This leaves open a number of possible remedies, if one wants to stay with the negative binomial model for the positive part. One could impose the restriction $\sigma_1^2 = 0$ (the Poisson case), or better perhaps, let $\sigma_1^2 = \sigma_2^2$. Alternatively, one may want to contemplate non-nested hurdle models, a class of models we consider in the following section.

6.3.4 Non-nested Hurdle Models

All models considered so far did nest a standard count data model, typically through a restriction of the sort $\theta_1 = \theta_2$. This was achieved since the hurdle process f_1 and the process for the positives, f_2, were based on the same distribution. If one sees the hurdle step as an altogether different process, one can as well use standard models for binary dependent variables at this

stage, such as the probit or the logit model. Such models were estimated by Grootendorst (1995). Gurmu (1998) discusses a model in which the hurdle step is parameterized by a generalized (asymmetric) logit model. Winkelmann (2004b) combines a probit hurdle model with a truncated-at-zero Poisson-log-normal model. The probability function of this probit-Poisson-log-normal model can be written as

$$P(y=0) = \Phi(x'\gamma) \tag{6.10}$$

$$P(y=k) = [1-\Phi(x'\gamma)] \times \int_{-\infty}^{\infty} \frac{\exp(-\exp(x'\beta+\varepsilon))\exp(x'\beta+\varepsilon)^y}{[1-\exp(-\exp(x'\beta+\varepsilon))]y!} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{\varepsilon}{\sigma}\right)^2} d\varepsilon \tag{6.11}$$

for $k = 1, 2, \ldots$, where Φ denotes the cumulative distribution function of the standard normal distribution. Note that the mixing in the second part of the distribution is done over the truncated Poisson model. As will be shown in the next chapter, this order is preferable over the alternative, to do the mixing first, and then form a truncated version of the resulting Poisson-log-normal model. Also, it is possible to generalize the model further by allowing correlation between the hurdle step and the process for the positives. A probit-Poisson-log-normal model with correlated hurdle is presented in Chap. 6.3.7.

While these hurdle models do not nest the Poisson or negative binomial model, they are also not strictly non-nested with the Poisson or negative binomial model, or among each other. In the terminology of Vuong (1989) these models are overlapping. The reason is that for certain parameter restrictions on both models, the two become equivalent. Take as an example the comparison between the logit-Poisson hurdle model and the simple Poisson model. In the logit-Poisson hurdle model

$$f_1(0; z, \gamma) = \frac{\exp(z'\gamma)}{1+\exp(z'\gamma)}$$

whereas in the simple Poisson model

$$f_2(0; x, \beta) = \exp(-\exp(x'\beta))$$

Now, assume that all slope parameters are set to zero and only non-zero intercepts, γ_0 and β_0 say, are left. The two probabilities, and with them the full distributions, are the same as long as as

$$\beta_0 = -\exp(\gamma_0) - \log(1-\exp(-\exp(\gamma_0)))$$

By the same technique, one can establish that the probit-Poisson-log-normal model and the Poisson model are overlapping, as are for instance the probit-Poisson-log-normal model and the standard negative binomial model.

In these cases, to discriminate between the two models following the methods proposed by Vuong (1989), one needs to follow the procedure for overlapping models. In general, this requires a pre-test before the usual statistic

is computed. However, in practice it is sufficient to establish that the condition for overlap, i.e., the restriction that all slope coefficients are zero, can be rejected in each model (see Vuong 1989, footnote 6).

6.3.5 Unobserved Heterogeneity in Hurdle Models

The standard approach to unobserved heterogeneity in hurdle models has introduced heterogeneity at the level of the parent distribution. Integration takes then place prior to the conversion into a hurdle model, the latter being based on modified densities $\tilde{f}_1$ and $\tilde{f}_2$ where

$$\tilde{f}_1 = \int f_1(y|u)g_1(u)du$$

and

$$\tilde{f}_2 = \int f_2(y|u)g_2(u)du$$

The hurdle negative binomial model is an example for this approach. In this case, f_1 and f_2 are Poisson probability functions, and g_1 and g_2 are gamma density functions.

The truncated probability function for the positives can then be written as

$$P(y|y > 0) = \frac{\tilde{f}_2(y)}{1 - \tilde{f}_2(0)} = \frac{\int f_2(y|u)g_2(u)du}{1 - \int f_2(0|u)g_2(u)du} \tag{6.12}$$

As pointed out by Santos Silva (2003), there is an alternative way of thinking of this problem, namely to define a distribution over the positive integers first (any truncated distribution will fulfill this requirement), and then do the mixing in a second step, over the positive part of the distribution only. In this case,

$$P(y|y > 0) = \int \frac{f_2(y|u)}{1 - f_2(0|u)}\tilde{g}_2(u)du \tag{6.13}$$

Clearly, these are not the same models, depending on how $g_2(u)$ and $\tilde{g}_2(u)$ are defined. In the regression context, where we consider conditional models, we typically assume independence between unobserved heterogeneity and the explanatory variables x. If we assume, for example, that $g_2(u|x)$ is a gamma density independent of x then (6.12) is a truncated at zero negative binomial distribution. If, however, f_2 is the Poisson distribution and $\tilde{g}_2(u|x)$ is a gamma density independent of x, the resulting probability function (6.13) is not that of a truncated negative binomial distribution. The question then becomes which of the two assumptions, and thus the two models, is more meaningful. Santos Silva (2003) argues that the population of interest is the actual population. In the case of the positive part of the hurdle models, this would favor an approach where the assumption is made that the unobservables in the (truncated) population of interest are independent of the x's, i.e., model

(6.13). In this way, one avoids the awkward step of needing to compute the integral $\int f_2(0|u)g_2(u)du$ although the zeros are generated by an altogether different process.

While hurdle models based on the negative binomial distribution assume that the unobservables are independent of the covariates in an hypothetical population, the probit-Poisson-log-normal model discussed above rather assumes that unobservables are independent of the covariates in the observed population. From this point of view, this makes it a more meaningful hurdle model.

6.3.6 Finite Mixture Versus Hurdle Models

A problem in health economics – how to model the demand for physician services – has prompted a controversy, whether finite mixture models or hurdle models are more appropriate for such data. The initial advocates of the hurdle model in this context, Pohlmeier und Ulrich (1995), maintained that the hurdle model may describe well the agency problem in the demand for doctor consultations, where the initial contact decision is made by the individual whereas further referrals are influenced by the physician's objectives. Arguably, then, two different parameterizations may be needed to capture this two-part decision process.

Proponents of finite mixture models take a less strict view. According to this view, every individual is a potential user but the population is composed of different types, or classes, of users. If there are two types, for instance, one could label them 'light users' and 'heavy users'. The econometrician does not observe which class an individual belongs to. Finite mixture models therefore are also called "latent class" models. The sample is a mixture of the two groups, and estimation of the group specific parameters and the group proportions is possible.

In a number of applications, both types of models have been estimated in order to determine which of the two better fits the data. For that purpose, one usually compares the hurdle negbin model

$$f(y) = f_{NB}(0;\theta_1)^d \left(\frac{1 - f_{NB}(0;\theta_1)}{1 - f_{NB}(0;\theta_2)} f_{NB}(y;\theta_2)\right)^{1-d}$$

where $d = 1 - \min(y, 1)$ as before, with a two component latent class negative binomial model

$$f(y) = \alpha f_{NB}(y;\theta_1) + (1-\alpha) f_{NB}(y;\theta_2)$$

The evidence is mixed. Using various model selection criteria (accounting for the fact that the second model has one additional parameter) Deb and Trivedi (2002) find that the finite mixture model is superior, although Winkelmann (2004b) shows that this is only the case for the Negbin hurdle models and not necessarily for the hurdle models as a class. Jimenéz-Martin, Labeaga and

Martinez-Granado (2002) report instances, where the hurdle model is better. See also Doorslaer, Koolman and Jones (2002). Bago d'Uva (2006) performs the obvious next step, by combining the two competing models and estimating a finite mixture negative binomial hurdle model, or "finite mixture of hurdle model", using panel data. Such a model is not very parsimonious, and identification problems loom large. Still, Bago d'Uva (2006) reports sensible estimates showing that, in this particular application to health care utilization, both the finite mixture part (here two classes) and the hurdle part offer statistically significant improvements over the more restrictive counterparts.

6.3.7 Correlated Hurdle Models

In the spirit of Chap. 5.2, it is straightforward to develop a generalized hurdle count data model where the hurdle process and the process for the positives are correlated. Such correlation may originate, for example, from common but unobserved variables. Such a model was proposed by Winkelmann (2004b).

In that paper, a probit model for the hurdle is combined with a truncated Poisson-log-normal model for the positives. Accordingly, it can be referred to as the *probit-Poisson-log-normal model*. Let z be a latent indicator variable such that

$$z = x'\gamma + \varepsilon$$

and

$$y = 0 \text{ iff } z \geq 0$$

Moreover, for the positive part of the distribution

$$y|y > 0 \sim \text{truncated Poisson}(\lambda)$$

where

$$\lambda = \exp(x'\beta + v)$$

The model is completed by the assumption that ε and v are bivariate normal distributed with mean 0 and covariance matrix

$$\Sigma = \begin{bmatrix} 1 & \rho\sigma \\ \rho\sigma & \sigma^2 \end{bmatrix}$$

To derive the log-likelihood function, note first that

$$\varepsilon|v \sim N(\rho v/\sigma, 1 - \rho^2)$$

and

$$\begin{aligned} P(y = 0|v) &= P(\varepsilon \geq -x'\gamma|v) \\ &= \Phi\left(\frac{x'\gamma + \rho v/\sigma}{\sqrt{1-\rho^2}}\right) \\ &= \Phi^*(v) \end{aligned}$$

Thus one obtains, with $d = 1 - \min(y, 1)$,

$$f(y|v) = \Phi^*(v)^d \times \left[(1 - \Phi^*(v))\frac{\exp(-\lambda(v))(\lambda(v))^y}{[1 - \exp(-\lambda(v))]y!}\right]^{1-d} \tag{6.14}$$

and

$$f(y) = \int_{-\infty}^{\infty} f(y|v)\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2}\left(\frac{v}{\sigma}\right)^2}\mathrm{d}v \tag{6.15}$$

The parameters β, γ and ρ can be estimated by maximum likelihood, using Gauss-Hermite integration to evaluate the likelihood function.

6.4 Zero-Inflated Count Data Models

6.4.1 Introduction

Zero-inflated Poisson or negative binomial models (ZIP, ZINB), like their hurdle-at-zero counterparts, address the problem that the data display a higher fraction of zeros, or non-occurrences, than is likely to be compatible with any fitted standard count regression model. The zero inflated model combines a binary variable c with a standard count variable y^* (with support over the nonnegative integers) such that the observed count y is given by

$$y = \begin{cases} 0 & \text{if } c = 1 \\ y^* & \text{if } c = 0 \end{cases} \tag{6.16}$$

If the probability that $c = 1$ is denoted by ω, the probability function of y can be written compactly as

$$f(y) = \omega d + (1 - \omega)g(y)\,, \qquad y = 0, 1, 2, \ldots \tag{6.17}$$

where $d = 1 - \min\{y, 1\}$ and $g(y)$ is a regular count data probability function such as the Poisson or the negative binomial probability function.

The difference between the zero-inflated model and the hurdle model is that in the latter, there is a single type of zeros whereas in the former one obtains two types of zeros: zero outcomes can either arise from regime 1 ($c = 1$) or from regime 2 ($c = 0$ and $y^* = 0$).

Which one of the two models – the zero-inflated model or the hurdle model – is more appropriate can be decided on statistical grounds, using methods developed for testing non-nested hypotheses, in particular the Vuong test. Such a test must account for the fact that the two models are overlapping (see Chap. 3.5.4), since the two models are equivalent if the slope coefficients are zero (Mullahy, 1986). The Poisson model is nested in the ZIP model for $\omega = 0$. Note, however, that the two models become non-nested if one adopts the specification $\omega = \exp(\theta)$ (or similar reparameterizations, where $\omega \neq 0$ for all finite parameter values).

Alternatively, the choice between hurdle and zero-inflated models can be made on substantive grounds as well. The question is then whether the characteristic assumption of zero-inflation models, namely two types of zero, is materially appealing or not. In the study of fertility, for example, we obtain a distinction between zero children due to infertility, in contrast to zero children due to choice. In this case, zero-inflation captures the difference between nature and choice. Another example is the study by Gameren and Woittiez (2002) where the count is the monthly number of subsidized home care days per month. Zeros can arise because a person is either not eligible, or does not know about the program, or alternatively, because an elderly person has simply no need. These are two competing explanations that can at least implicitly be distinguished by applying this type of model.

In the analysis of the determinants of R&D productivity, where output is frequently measured in terms of patent applications, zeros can arise in two different ways as well. First, there are firms that have decided not ever to apply for a patent, regardless of whether an invention was made. Secondly, there are firms that register patents in practice, but not necessarily in a given period if no invention was made. Lambert (1992) has referred to the first type of zeros as being *strategic*, whereas the second type is incidental.

Zero inflated models have become quite popular in the recent applied count data literature, and they appear to be more frequently used than hurdle models. Economic applications of zero-inflated models are often based on the zero-inflated negative binomial model. Examples include Grootendorst (1995) on prescription drug utilization, List (2002) on the number of job interviews secured by a job seeker, and Tomlin (2000) on the empirical connection between exchange rates and the number of foreign direct investment activities. Beckmann (2002) uses the zero-inflated negative binomial model for modeling the number of apprentices trained by a firm, and Kahn (2005) applies it to the number of deaths from natural disasters.

6.4.2 Zero-Inflated Poisson Model

Mullahy (1986) discusses the zero-inflated Poisson model with constant ω. Lambert (1992) extends it by specifying a logit model for ω in order to capture the influence of covariates on the probability of extra zeros:

$$\omega = \frac{\exp(z'\gamma)}{1 + \exp(z'\gamma)} \tag{6.18}$$

The zero-inflated Poisson model has mean function

$$\mathrm{E}(y|x, z) = (1 - \omega)\exp(x'\beta) = \frac{\exp(x'\beta)}{1 + \exp(z'\gamma)} \tag{6.19}$$

Hence, the Poisson pseudo maximum likelihood estimator ceases to be consistent if the zero inflated model is the true data generating process, unless x and z are independent. In this case, only the constant is affected, and we can write

$$\mathrm{E}(y|x) = \exp(\tilde{\beta}_0 + x'\beta)$$

where $\tilde{\beta}_0 = \beta_0 + \log \mathrm{E}[1/(1 + \exp(z'\gamma)]$. In general, though, x and z are not independent, and it is important to account for this type of departure from the Poisson model.
The log-likelihood function of the zero-inflated Poisson model is

$$\ell(\gamma, \beta) = \sum_{y_i=0} \log(\exp(z_i'\gamma) + \exp(-\exp(x_i'\beta)) + \sum_{y_i>0} y_i x_i'\beta - \exp(x_i'\beta) - \log(y!) - \sum_{i=1}^{n} \log(1 + \exp(z_i'\gamma)) \quad (6.20)$$

In contrast to the hurdle Poisson model presented in Chapter 6.3, the log-likelihood function cannot be separated in two functions of β and γ, respectively. Although this makes estimation somewhat more complex, analytical expressions for the first and second derivatives of the log-likelihood function can be easily derived, and estimation therefore is not difficult after all. Often, exclusion restrictions are used in empirical work, such that neither x nor z are proper subsets of each other.

A number of variants of the zero-inflated Poisson model have been discussed in the literature:

- In *zero-altered* models (see also Chapt 2.4.3), ω is not restricted to be positive. The requirement $f(0) > 0$ implies that $\omega > -g(0)/(1 - g(0))$. For $0 > \omega > -g(0)/(1 - g(0))$, we obtain a model with zero-deflation. Of course, ω does not represent a probability in this case, which means that we loose the selection interpretation (6.16), and we cannot employ the logit specification (6.18) either.
- One can inflate (or alter) the probabilities for other outcomes as well. As an example, Melkersson and Roth (2000) have considered a zero-and-two inflated count data model in an analysis of the demand for children among Swedish couples. They wanted to test whether the outcome "two" occurs more often than predicted under a standard count data model, indicating the presence of a norm for an "ideal" family size.
- A model with correlated unobserved heterogeneity in both index functions has been considered by Crépon and Duguet (1997b). In their case, the process for the extra zeros is a probit model, and the count process is a Poisson-log-normal model. Further details can be found in Chap. 5.2.5.

Alternative Estimation

Santos Silva and Covas (2000) have pointed out that the conditional distribution $f(y|y > 0)$ does not depend on ω:

$$f(y|y > 0) = \frac{(1 - \omega)g(y)}{1 - [\omega + (1 - \omega)g(0)]}$$

$$= \frac{g(y)}{1 - g(0)} \quad y = 1, 2, \ldots \tag{6.21}$$

Hence, one can use positive observations only and estimate the parameters of g using a truncated-at-zero count data model without the need to specify ω. In fact, this is the same estimator as the second part of the hurdle model. The advantage is that such an estimator is robust with respect to a misspecification of ω, and is can serve as a basis for specification tests.

6.4.3 Zero-Inflated Negative Binomial Model

The extension from a zero-inflated Poisson (ZIP) to a zero inflated negative binomial model (ZINB) is straightforward. For example, with a Negbin II specification for the count data part, and a logit specification for the extra-zeros, the log-likelihood function is given by

$$\begin{aligned}\ell(\gamma, \beta, \alpha) = &\sum_{y_i=0} \log(\exp(z_i'\gamma) + \alpha(\log\alpha - \log(\exp(x_i'\beta) + \alpha)) \\ &+ \sum_{y_i>0} \log(\Gamma(\alpha + y_i)/\Gamma(\alpha) + \alpha\log(\alpha - \log(\exp(x_i'\beta)) \\ &+ y_i(x_i'\beta - \log(\exp(x_i'\beta) + \alpha)) - \sum_{i=1}^{n} \log(1 + \exp(z_i'\gamma))\end{aligned}$$

The ZINB model nests the ZIP model, which means that the Poisson restriction can be tested using a simple likelihood-ratio or Wald test.

6.4.4 Marginal Effets

As always in such multi-index models there are several marginal effects of interest. One is the overall marginal mean effect, $\partial \mathrm{E}(y|x)/\partial x_j$. This effect can be decomposed into the effect at the extensive margin, $\partial P(y > 0|x)/\partial x_j$, and the effect at the intensive margin, $\partial \mathrm{E}(y|y > 0x)/\partial x_j$. Recall that, for the zero inflated Poisson model, we have

$$\mathrm{E}(y|x, z) = (1 - \omega)\exp(x'\beta)$$

and

$$P(y = 0|x, z) = \omega + (1 - \omega)\exp(-\exp(x'\beta))$$

where ω is modeled as a logistic function of z, i.e.,

$$\omega = \frac{\exp(z'\gamma)}{1 + \exp(z'\gamma)}$$

Moreover, from (6.21),

$$\mathrm{E}(y|y>0,x) = \frac{\exp(x'\beta)}{1-\exp(-\exp(x'\beta))}$$

We see that $\mathrm{E}(y|x)$ and $P(y=0|x)$ are both functions of x and z. As a consequence, marginal effects depend on whether or not z and x overlap. We consider in the following marginal effects with respect to a variable that appears in both parts of the model, such that $x_j = z_k$ for some j,k. With this assumption

$$\begin{aligned}\frac{\partial \mathrm{E}(y|x,z)}{\partial x_j} &= -\frac{\partial \omega}{\partial x_j}\exp(x'\beta) + (1-\omega)\frac{\partial \exp(x'\beta)}{\partial x_j} \\ &= -\frac{\exp(z'\gamma)\exp(x'\beta)}{(1+\exp(z'\gamma))^2}\gamma_k + \frac{\exp(x'\beta)}{1+\exp(z'\gamma)}\beta_j\end{aligned}$$

This marginal effect is not necessarily positive. For example, if $\gamma_k > 0$ and $\beta_j > 0$, there are two effects working in opposite direction: an increase in x_j increases the probability of an extra zero; at the same time the mean of the base model is increased as well. The overall effect is then ambiguous. In this case, a necessary and sufficient condition for a positive marginal mean effect is that

$$\frac{\beta_j}{\gamma_k} > \frac{\exp(z'\gamma)}{1+\exp(z'\gamma)} = \omega$$

As usual, one needs to evaluate these marginal mean effects at some appropriate value of the explanatory variables. Alternatively, one may compute marginal effects at the extensive and intensive margins, as outlined in 6.2.5.

As to the effects at the extensive margin, we obtain

$$\begin{aligned}\frac{\partial P(y=0|x)}{\partial x_j} &= \frac{\exp(z'\gamma)}{(1+\exp(z'\gamma))^2}(1-\exp(-\exp(x'\beta))\gamma_k \\ &\quad - \frac{\exp(x'\beta)}{1+\exp(z'\gamma)}(1-\exp(-\exp(x'\beta))\beta_j\end{aligned}$$

Again, these effects can be either positive or negative, depending on the relative magnitudes of γ_k and β_j. Finally, using (6.21),

$$\frac{\partial \mathrm{E}(y|y>0,x)}{\partial x_j} = \left(\frac{\exp(x'\beta)}{1-\exp(-\exp(x'\beta))} - \frac{\exp(x'\beta)^2\exp(-\exp(x'\beta))}{(1-\exp(-\exp(x'\beta)))^2}\right)\beta_j$$

The term in parentheses is positive, since it can be re-written as the ratio of $1-P(y=0|x)-P(y=1|x)$ and $1-P(y=0|x)$. As a consequence, the sign of the marginal effect at the intensive margin is equal to that of β_j.

6.5 Compound Count Data Models

The concept of compounding, and its application to count data distributions, has been introduced in Chap. 2.5.2. To repeat the essential idea, a count

variable Z has a compound count data distribution if it can be written as a random sum

$$Z = \sum_{i=1}^{N} X_i$$

where N is the number of summands, $X_i, i = 1, \ldots, N$ are identically and independently distributed discrete random variables, and either N, or X_i, or both, have support over the non-negative integers. Compound distributions are sometimes also referred to as "stopped-sum distribution".

The appeal of this framework is that it provides a very general approach for building multi-index models since covariates can be introduced separately in the two parts of the model, the "N"-part and the "X"-part. In fact, both zero-inflated and hurdle count data models are compound models (where N is a binary 0/1 variable). By making different assumptions for N and X, alternative models can be obtained. We present here two such generalized models, the multi-episode model and the Poisson model with underreporting.

6.5.1 Multi-Episode Models

A multi-episode model is discussed by Santos Silva and Windmeijer (2001). They motivate their approach in an application to the demand for health services, as measured by the number of doctor visits during a given period of time. They relate the underlying demand process to illness episodes that each generate a certain number of visits to the general practitioner or to specialists ("referrals"). The total number of visits Z is then equal to the sum of visits in each episode X_i over the N illness spells. They assume that N has a Poisson distribution, whereas X has a logarithmic distribution. The logarithmic distribution has support $1, 2, \ldots$, which makes sense, since an illness episode includes by definition at least one visit. Conveniently, this particular compound distribution is of a negative binomial form (see Chap. 2.5.2). Further details of their model are omitted here, as they are discussed together with the other negative binomial models in Chap. 4.3.4 rather than here.

6.5.2 Underreporting

Consider the following data generating process: Events occur randomly over time according to a Poisson process. In contrast to the standard Poisson model, however, only a subset of events is reported. The number of reported events is smaller than the total number of events. "Reporting" can be understood in a very broad sense. The basic distinction is whether events are "successful", or "unsuccessful". Observed counts give the number of successful events.

This model is applicable in many situations: In labor economics it can describe the frequency of job changes in a given period (Winkelmann and

Zimmermann, 1993c, Winkelmann, 1996b). This frequency will depend on both the frequency at which outside job offers are received, and the probability that outside offers are accepted. In industrial organization, the model can be used for the number of firms entering an industry in a given period (Berglund and Brännäs, 1995). Here, the base entity is the number of potential entrants who decide whether or not to enter. Finally, in modeling accident proneness one might distinguish between the total number of accidents (in a region, for instance) and the number of accidents involving fatalities.

Poisson-Logistic Regression

In the simplest case, the outcome of interest is Poisson distributed, and the reporting process is stationary and independent of the Poisson variable. The distinction between reported (or "successful") and unreported (or "unsuccessful) events is assumed to follow an independent binomial distribution. The model is then in the form of a compound Poisson distribution, and the total number of successful events is again Poisson distributed with a modified mean function. (The strong independence assumption is relaxed in Chap. 6.5.4)

Formally, let the total number of events y^* in a given period of time be Poisson distributed with

$$f(y^*) = \exp(-\lambda)\lambda^{y^*}/y^*! \tag{6.22}$$

where

$$\lambda = \exp(z'\gamma) \tag{6.23}$$

and z is a vector of individual covariates. Also assume, following Winkelmann and Zimmermann (1993c), that the binomial model is of the logistic form

$$P(\text{Event is successful}) = \Lambda(x'\beta) \tag{6.24}$$

where x are individual covariates and Λ is the *logistic* cumulative

$$\Lambda(x'\beta) = \frac{\exp(x'\beta)}{1+\exp(x'\beta)} \tag{6.25}$$

The set of explanatory variables z and x can be overlapping, but they may not be identical. Thus, some exclusion restrictions are required. It will be assumed that consecutive decisions determining the success or non-success of an event are independent and identical. The number of successful events y is then given by the sum of i.i.d. distributed Bernoulli variables, where the number of summands is itself a random variable:

$$y = B_1 + \ldots + B_{y^*} \tag{6.26}$$

where $B \sim Bn(1, \Lambda(x'\beta))$ and $y^* \sim \text{Poisson}(\exp(z'\gamma))$. Under independence, the number of successful events y_i has a Poisson distribution with parameter

$$\lambda^S = \exp(z'\gamma)\Lambda(x'\beta) \tag{6.27}$$

where z is a $(k_1 \times 1)$ vector and x is a $(k_2 \times 1)$ vector (see Chap. 2.5.2 and Feller, 1969). Estimation of the parameters β and γ by maximum likelihood is straightforward. The log-likelihood function has the form

$$\ell(\beta,\gamma) = \sum_{i=1}^{n} -\frac{\exp(x_i'\beta + z_i'\gamma)}{1+\exp(x_i'\beta)} + y_i \log\left[\frac{\exp(x_i'\beta + z_i'\gamma)}{1+\exp(x_i'\beta)}\right] - \log y_i! \tag{6.28}$$

Collecting the coefficients β and γ in a parameter vector θ, the score vector can be written as

$$\frac{\partial \ell(\theta; y, z, x)}{\partial \theta} = \sum_{i=1}^{n} \left(\frac{y_i - \lambda_i^S}{\lambda_i^S}\right) \frac{\partial \lambda_i^S}{\partial \theta} \tag{6.29}$$

or

$$\frac{\partial \ell(\theta; y, z, x)}{\partial \theta} = \sum_{i=1}^{n} (y_i - \lambda_i^S) \begin{bmatrix} z_i' \\ x_i'(1 - \Lambda(x_i'\beta)) \end{bmatrix} \tag{6.30}$$

If z contains an intercept, (6.30) states that the sum of the residuals $u_i = y_i - \lambda_i^S$ is equal to zero. The Hessian matrix has the form

$$\frac{\partial^2 \ell(\theta; y, z, x)}{\partial \theta \partial \theta'} = \sum_{i=1}^{n} -\lambda_i^S \times$$

$$\begin{bmatrix} z_i z_i' & z_i x_i'(1 - \Lambda(x_i'\beta)) \\ \cdot & x_i x_i' \left[(1 - \Lambda(x_i'\beta))^2 + \dfrac{(y - \lambda_i^S)}{\lambda_i^S} \Lambda(x_i'\beta)(1 - \Lambda(x_i'\beta))\right] \end{bmatrix} \tag{6.31}$$

Hence, the information matrix of the sample is given by

$$I_n(\theta) = \sum_{i=1}^{n} \lambda_i^S \begin{bmatrix} z_i z_i' & z_i x_i'(1 - \Lambda(x_i'\beta)) \\ x_i z_i'(1 - \Lambda(x_i'\beta)) & x_i x_i'(1 - \Lambda(x_i'\beta))^2 \end{bmatrix} \tag{6.32}$$

Identification of the parameters requires that $I_n(\theta)$ is nonsingular for arbitrary values in the parameter space. For instance, the information matrix is singular for $x = z$ at $\beta = 0$ (which implies that $1 - \Lambda(x_i'\beta)|_{\beta=0} = 0.5$.

It can be shown more generally that identifiability requires that neither x nor z does belong to the column space of the other. Under the assumptions of the model, the maximum likelihood estimator $\hat{\theta}$ is consistent for θ and $\sqrt{n}(\hat{\theta} - \theta) \to N(0, I(\theta)^{-1})$, where $I(\theta)^{-1}$ is the inverse of the information matrix of an observation.

6.5.3 Count Amount Model

In the Poisson-logistic model, the probability of reporting an event is modeled as a function of individual specific (but event-unspecific) covariates. In a related model, the probability of recording an event depends on the (partially observed) "magnitude" of the event. The structure of this model is best illustrated with an application to consumer purchase behavior for which it was originally developed by van Praag and Vermeulen (1993). In this application, purchases are reported, together with their money value, as long as they exceed a minimum threshold value a. Let y^* denote the total number of purchases over a given period of time, and let y denote the number of recorded purchase occurrences. Denote the purchase amount by c^*. A purchase is recorded whenever $c^* > a$. The probability for this event is given by $P(c^* > a) = 1 - F(a)$, where F is the cumulative density function of c^*. In this set-up, observed purchase values c have a truncated density function $f(c)/(1 - F(a))$.

The model is completed by imposing some distributional assumptions. First, c^* is assumed to be normally distributed, conditional on individual specific covariates: $c^* = z\gamma + \varepsilon$ where ε has a normal distribution $N(0, \sigma^2)$. Second, y^* is assumed to be Poisson (or negative binomial) distributed with parameter $\lambda = \exp(x'\beta)$.

Since both the number of events and the money values are observed, van Praag and Vermeulen estimate the parameter vector $\theta = [\beta, \gamma]$ from the joint likelihood function of $c = (c_1, \ldots, c_y)$ and y which is given by

$$
\begin{aligned}
g(y, c) &= \prod_{i=1}^{y} \frac{f(c; \gamma)}{1 - F(a)} \\
&\quad \times \sum_{y^*=y}^{\infty} f(y^*; \beta) \frac{y^*!}{y!(y^* - y)!} (1 - F(a))^y (F(a))^{y^* - y} \\
&= \prod_{i=1}^{y} f(c; \gamma) \sum_{y^*=y}^{\infty} f(y^*; \beta) \frac{y^*!}{y!(y^* - y)!} (F(a))^{y^* - y}
\end{aligned}
\tag{6.33}
$$

where $f(y^*; \beta)$ is a Poisson or negative binomial probability function and $f(c; \gamma)$ is the normal density. The term under the summation sign is a Poisson-Binomial mixture. To understand the meaning of this expression, note that the event "y purchases are recorded" can arise in a multitude of ways:

1. there were $y^* = y$ purchases, all with amounts greater than a and therefore all recorded.
2. there were $y^* = y + 1$ purchases, of which y with amounts greater than a and one with amount less than a (and therefore unrecorded).
3. there were $y^* = y + 2$ purchases, of which y with amounts greater than a and two with amount less than a (and therefore unrecorded), and so forth.

The probability for each event in the above list is a joint probability $f(y^*, y)$ which can be written as a product of marginal and conditional distribution:

$$f(y^*, y) = f(y^*)f(y|y^*) \tag{6.34}$$

The first probability function on the right side is a Poisson (or negative binomial) distribution. Under the assumption that purchase amounts at subsequent purchase occasions are independent, the second (conditional) probability function is a binomial distribution with parameter $p = 1 - F(a)$, i.e. the probability of success, here recording an event, is equal to the probability that the purchase amount exceeds the threshold of a. Without the independence assumption, it would not be possible to obtain such simple probability expressions.

As van Praag and Vermeulen (1993) point out this model has a wide range of potential applications. Examples include the modeling of the number of insurance claims where the insurance includes a deductible amount, the modeling of crime statistics, where official authorities do not file formal reports for minor crimes, or the number of unemployment spells, where only spells exceeding a certain minimum duration are observed.

6.5.4 Endogenous Underreporting

A count data model with endogenous reporting was considered in Winkelmann (1997, 1998). The model is closely related to the models on incidental censoring and truncation of Chap. 5.2. A restriction of the standard model is the assumption of independence between the count process and the binary reporting outcome. Consider, for instance, the study by Winkelmann and Zimmermann (1993c), where the model is applied to data on labor mobility. y^* gives then the (unobserved) number of job offers, $\lambda = \exp(x'\beta)$ the offer arrival rate, p the acceptance probability and y the (observed) number of accepted offers. The explicit assumption is that

a) the offer arrival rate is a deterministic function of observed covariates, and
b) the offer arrival rate is independent of the acceptance probability.

Yet, it is unreasonable to assume that all relevant variables are observed in practice and that arrival rates and acceptance decisions are independent. For instance, economic models of efficient job search predict that the reservation wage depends on the offer arrival rate and hence a correlation between the two should exist (See Mortensen, 1986). Therefore, a more general model that allows for endogenous underreporting is desirable. Such a model is now introduced.

Let $y^*|v$ have a count data distribution with mean

$$\mathrm{E}(y^*|x, v) = \exp(x'\beta + v) \tag{6.35}$$

As before, an event j is reported and $c_j = 1$ if the net utility from doing so is positive, i.e.

$$c_j^* = c^* = z'\gamma + \varepsilon > 0 \tag{6.36}$$

where, by assumption, the net utility does not depend on the specific event. Furthermore, assume that v and ε are jointly normal distributed with correlation ρ. Note, that this model is based on a probit-type specification whereas the standard underreporting model was based on the logit model. This change is dictated by convenience as the probit model leads to a straightforward extension for the correlated case.

The number of reported counts is given by

$$y = \sum_{j=1}^{y^*} c \tag{6.37}$$

To derive the probability function of y, consider first the case where v is given. As before

$$P(c = 1|v) = \Phi^*(v, z) \tag{6.38}$$

where $\Phi^*(v, z)$ is defined as in (5.32). Moreover, conditional on v, x and z, c and y^* are independent. Assume that $y^*|v$ is Poisson distributed. It follows directly from results in Chap. 6.5.2 that the conditional distribution of the reported number of events, y, is also a Poisson distribution with mean

$$\tilde{\lambda} = \exp(x'\beta + v) \times \Phi^*(v) \tag{6.39}$$

while $y|x, z$ has distribution

$$g(y|x, z) = \int_{-\infty}^{\infty} \frac{\exp(-\tilde{\lambda}(v))\tilde{\lambda}(v)^y}{y!} f_v(v)\mathrm{d}v \tag{6.40}$$

or, in explicit notation

$$g(y|x, z; \beta, \gamma, \rho, \sigma) = \tag{6.41}$$

$$\int_{-\infty}^{\infty} \exp\left[-\exp(x'\beta + v)\Phi\left(\frac{z'\gamma + \rho v/\sigma}{\sqrt{1-\rho^2}}\right)\right]$$

$$\times \left[\exp(x'\beta + v)\Phi\left(\frac{z'\gamma + \rho v/\sigma}{\sqrt{1-\rho^2}}\right)\right]^y \times \frac{1}{y!\sigma}\phi(v/\sigma)\mathrm{d}v$$

The parameters of the model, β, γ, ρ, and σ can be estimated by maximum likelihood. The resulting log-likelihood function involves simple integrals that can be evaluated by Gauss-Hermite quadrature. Details are given in Appendix B.

The model is quite general and encompasses a variety of interesting special cases that can be tested using parametric restrictions. For $\rho = 0$ the selection and count equations are independent. For $\rho = 0$ and $\sigma = 0$, the model reduces to a version of the Poisson-logistic regression model in Winkelmann and

Zimmermann (1993c) where the logit type expression for the reporting probability is replaced by a probit type expression. Positive values for σ indicate unobserved heterogeneity in the count regression. In particular, the implicit variance function for y^* is

$$\text{Var}(y^*|x) = \lambda + \alpha\lambda^2 \tag{6.42}$$

where $\alpha = \exp(2\sigma_u^2) - \exp(\sigma_u^2)$.

6.6 Quantile Regression for Count Data

Quantile regression has been in use in the context of linear regression models for a long time (see Koenker and Hallok, 2001, for a general introduction to the topic). Regression based on least absolute deviations, or median regression, offers an alternative to least squares. One of the advantages of quantile regression is that is robust to outliers. More importantly, though, quantile regressions can be performed for arbitrary quantiles of a distribution. Seen in this way, quantile regression becomes a tool for modeling the effect of regressors on the full distribution of the outcome variable, rather than modeling their effect on the first noncentral moment of the distribution only.

Quantile regression is not the only possibility to model the whole distribution. In fact, it has been pointed out before that all count data models proper are probability models and allow inferences to be drawn on all possible aspects of the outcome distribution, including the computation of marginal probability effects. Also, if the underlying model is specified with sufficient flexibility (for instance using hurdle type models) such inferences are not tautological but informed by data. Probability based models and quantile based models are two sides of the same coin: probability models are based on the representation of a random variable through its probability function, whereas quantile models are based on the representation of a random variable through its distribution function.

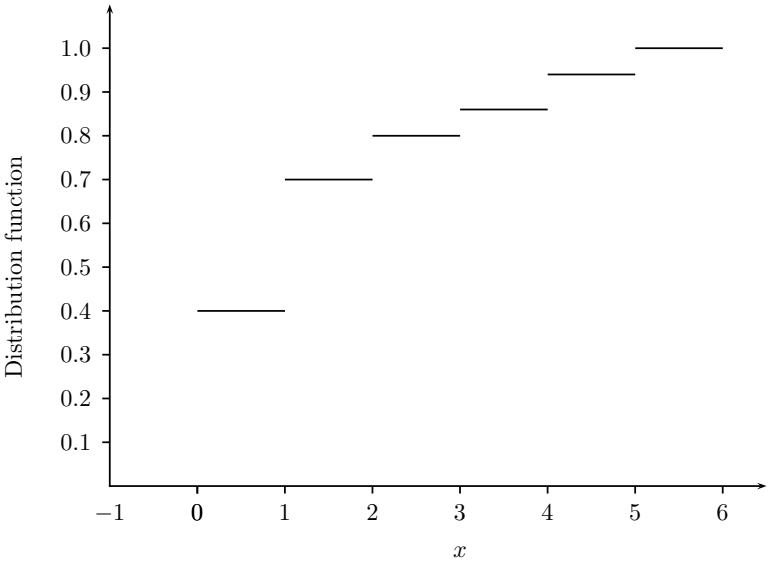
The main drawback of the latter is then of a more technical nature, namely the fact that the distibution function of a discrete random variable is not continuous. See Fig. 6.2 for a stylised distribution function with its typical jumps. However, this difficulty can be overcome, as shown by Machado and Santos Silva (2005). Let y be a count random variable. The α-quantile of y is defined by

$$Q_y(\alpha) = \min(\eta|P(y \leq \eta) \geq \alpha)$$

where $0 \leq \alpha < 1$. The object of interest is the conditional quantile $Q_y(\alpha|x)$. Since $Q_y(\alpha|x)$ has the same support as y, it is discrete and cannot be a continuous function of x (such as $\exp(x'\beta)$). Therefore, Machado and Santos Silva suggest to introduce "jittering": consider a new variable z, obtained by adding a uniform random variable to the count variable

$$z = y + u \text{ where } u \sim \text{ uniform } [0, 1)$$

Fig. 6.2. Count Data Distribution Function Without Uniform Distribution Added



where y and u are independent. Hence, z has density function

$$f(z) = \begin{cases} p_0 \text{ for } 0 \leq z < 1 \\ p_1 \text{ for } 1 \leq z < 2 \\ \text{and so forth} \end{cases}$$

(using notation $P(Y = k) = p_k$). Moreover, the distribution function of z can be written as

$$F(z) = \begin{cases} p_0 z \text{ for } 0 \leq z < 1 \\ p_0 + p_1(z-1) \text{ for } 1 \leq z < 2 \\ \text{and so forth} \end{cases}$$

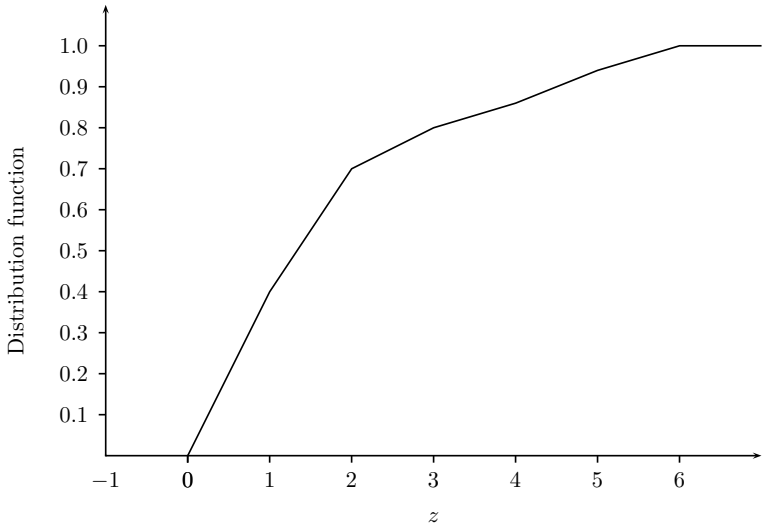
The distribution function of z is also shown in Fig. 6.3. We see that the quantiles of z are continuous. For example,

$$z_\alpha = \frac{\alpha}{p_0} \text{ for } \alpha < p_0$$

$$z_\alpha = 1 + \frac{\alpha - p_0}{p_1} \text{ for } p_0 \leq \alpha < p_0 + p_1$$

Having established the relationship between the probability function of y and the quantiles of z, we can now turn to the estimation of the quantiles. The probabilities being unknown, the conditional z-quantiles are estimated directly

Fig. 6.3. Count Data Distribution Function With Uniform Distribution Added



by adopting a regression framework. Machado and Santos Silva (2005) specify the quantiles $Q_z(\alpha|x)$ as

$$Q_z(\alpha|x) = \alpha + \exp(x'\gamma(\alpha)), \quad \alpha \in (0,1) \tag{6.43}$$

The reason for adding α on the right side of (6.43) is that the z-quantiles are bounded from below at α. The minimum is reached in the degenerate case where $p_0 = 1$. The dependent variable z (and with it the z-quantiles since quantiles are invariant under transformation, i.e., $Q_{g(z)} = g(Q_z)$) can be transformed such that the transformed quantile function is linear in x and γ. Observe that

$$Q_{T(z;\alpha)}(\alpha|x) = x'\gamma(\alpha)$$

where

$$T(z;\alpha) = \begin{cases} \log(z-\alpha) & \text{for } z > \alpha \\ \log(\xi) & \text{for } z \leq \alpha \end{cases} \tag{6.44}$$

and $0 < \xi < \alpha$.

The model suggest the following empirical implementation. First, one adds uniformly distributed pseudo random numbers to the observed counts. Second, one transforms the resulting data according to (6.44). Third and finally, the parameter estimates are obtained as solution to

$$\min \sum_{i=1}^{n} \rho_\alpha(T(z_i;\alpha) - x_i'\gamma)$$

where $\rho_\alpha(\nu) = \nu \times (\alpha - I(\nu < 0))$. For example, if $\alpha = 0.5$, the *rho* function returns simply the absolute value $0.5|T(z_i; \alpha) - x_i'\gamma|$.

Machado and Santos Silva (2005) prove consistency and asymptotic normality of this estimator. Although the quantile function is not differentiable everywhere (the distribution function has corners), these points do not affect the derivation as long as there is at least one continuous regressor, because in this case these corner points have measure zero.

The final question is how the parameters should be interpreted. In this approach, one estimates $Q_z(\alpha|x)$. The object of interest is $Q_y(\alpha|x)$, though. There is a correspondence between the two quantile functions, since $Q_y(\alpha|x) = \text{int}^*[\text{Q}_\text{Z}(\alpha|\text{x})]$, where $\text{int}^*(a)$ is the ceiling function which returns the smallest integer greater than, or equal, to a. Hence, for example, testing the null hypothesis that a variable has no effect on $Q_Y(\alpha|x)$ is equivalent to testing that a variable has no effect on $Q_Z(\alpha|x)$. Recent applications of this method are Winkelmann (2006) to the number of physician visits, and Miranda (2008) to the number of children.

7

Correlated Count Data

7.1 Multivariate Count Data

Multivariate count data are likely to have a non-trivial correlation structure. For instance, omitted variables may simultaneously affect more than one count. The modeling of the correlation structure is important for the efficiency of the estimator and the computation of correct standard errors, i.e., valid inference. Beyond that, the nature of the stochastic interaction between several counts may be of independent and intrinsic interest.

Following the usual notation for multivariate data let y_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, J$ denote the count for individual i and outcome j. Let $y_i = (y_{i1} \cdots y_{iJ})'$ denote the vector of counts for individual i over the different outcomes. Interest is in a class of models where observations are uncorrelated across individuals but correlated over outcomes:

$$\mathrm{Cov}(y_{ij}, y_{kl}) \begin{cases} = 0 & \text{for } i \neq k \\ \neq 0 & \text{for } i = k,\ j \neq l \end{cases}$$

In matrix notation, correlated outcomes mean that $\mathrm{Var}(y_i)$ is not a diagonal matrix. Five models for correlated counts are discussed in this chapter

- Multivariate Poisson Model
- Multivariate Negative Binomial Model
- Multivariate Poisson-Gamma Mixture Model
- Multivariate Poisson-Log-Normal Model
- Latent Poisson-Normal Model

The first three models have in common that they are based on a so-called one-factor structure: correlation is generated through an individual specific random factor u_i that does not vary over outcomes. A direct implication of this relatively simplistic assumption is that the covariance structure of these models is restricted to non-negative correlations. A more flexible covariance structure is provided by the multivariate Poisson-log-normal and the latent Poisson-normal models, two examples of multi-factor models. The chapter

closes with a discussion of copulas, a general method for constructing multivariate distributions for given marginals, although applications of such methods for count data modeling have just started.

The class of models pursued in this chapter encompasses a variety of interesting sub-cases and is more general than might appear at first. The initial interpretation is in terms of genuine multivariate outcomes. One example occurs when studying the provision of health services where a researcher might be interested in a joint analysis of the number of individual visits to a doctor and to a non-doctor health specialist (Gurmu and Elder, 1998). A further example is in the study of labor mobility, where one may be interested in the joint determination of voluntary and involuntary job changes (Jung and Winkelmann, 1993).

An alternative interpretation is in terms of seemingly unrelated regression (SURE). In this case, one considers a single outcome variable, for example the annual number of airtraffic incidents by major air carrier over a 30 year period as in Rose (1990). In a sense, these are also panel data, see below. However, they differ from usual panel data in at least two respects. First, in typical SURE applications, the time dimension is large relatively to the cross-section dimension. As a consequence, it becomes possible to estimate unit-specific constants and slope parameters for each cross-sectional unit. Second, because the number of cross sectional units is small and these units may interact (such as firms operating in the same market, or countries) and the random sampling assumption may be invalid. To be specific, under the assumption of the SURE model, observations are assumed to be (contemporaneously) correlated across observations (and hence not an independent random sample) but uncorrelated over time, given the unit-specific fixed effects. In order to make the SURE notation conform to the multivariate approach, let i denote time and j denote the cross-sectional observations. In this way, $\mathrm{Cov}(y_{ij}, y_{kl})$ is a contemporaneous covariance for $i = k$, $j \neq l$.

A third area of application is to panel data. As mentioned before, both panel and SURE-type data have an individual and a time dimension. However, for panel data proper, the individual dimension is typically large relative to the time dimension. Often, such datasets includes measurements for thousands of individuals over a few years. Moreover, contemporaneous correlation is precluded by the assumption of random sampling, a reasonable assumption in such context. In this case, the interest is rather in the correlation of observations for a given individual over time. Such correlation may arise, for instance, from individual specific random coefficients. To adjust the multivariate set-up to the panel case i denotes the individual and j the period. Models for panel data are discussed in greater detail in Chap. 7.2.

A further difference between the panel view and the other two interpretation is that for multivariate and SURE data, one is interested in the correlation structure mainly in order to improve efficiency. For panel data, however, omitted variable bias, and hence the question whether the individual specific effects should be treated as fixed or random, is a major issue. Because T is

small relative to n, estimating n separate constants is not a trivial issue. It reduces the degrees of freedom considerably. Typical multivariate and SURE data do not have this problem, and in these case, we can always let

$$\mathrm{E}(y_{ij}|x_{ij}) = \exp(x'_{ij}\beta_j)$$

Random coefficients models, though feasible (see Chib, Greenberg and Winkelmann, 1998, and Chap. 8.4) have been rarely used in practice so far. In order to keep notation simple, the following exposition does abstract from heterogeneity in slope parameters, be it observed or unobserved.

The classification of data into multivariate, SURE, or panel may be useful. It should not hide the fact, however, that the labels sometimes will be ambiguous. Also, hybrid cases clearly can exists, such as multivariate panel data, as discussed in Million (1998) and in Riphahn, Wambach and Million (2003).

7.1.1 Multivariate Poisson Distribution

In this chapter, the multivariate Poisson (MVP) distribution is derived and characterized. Its appeal stems both from its relative simple structure and its derivation based on a common additive error. The three main features of the multivariate Poisson distribution are

1. Two MVP distributed random variables are independent if and only if they are uncorrelated.
2. The marginal distributions of the MVP distribution are univariate Poisson.
3. The conditional distributions of the bivariate Poisson distribution are sums of an independent Poisson and binomial distribution. The conditional expectation functions are linear.

Kocherlakota and Kocherlakota (1992) provide a comprehensive discussion of the MVP distribution. They also point out that the adopted formulation is just one possible extension of the Poisson distribution to the multivariate setting (See Lakshminarayana, Pandit and Rao, 1999, for a recent example for a bivariate Poisson distribution based on a polynomial factor). So far, the MVP distribution seems to be the only one that has been put to practical use in econometrics, and we thus restrict the attention to this particular model.

The derivation, sometimes referred to as "trivariate reduction method", is based on a convolution structure: Let the random variables z_{ij}, $j = 1, \ldots, J$ and u_i have independent Poisson distributions with $z_{ij} \sim \text{Poisson}(\lambda_{ij})$, and $u_i \sim \text{Poisson}(\gamma)$ where $\lambda_{ij} = \exp(x'_{ij}\beta)$. New random variables y_{ij} can be constructed as

$$y_{ij} = z_{ij} + u_i \tag{7.1}$$

Using the convolution property of independent Poisson variables, $y_{ij} \sim \text{Poisson}(\lambda_{ij} + \gamma)$, $j = 1, \ldots, J$, that is, the bivariate Poisson distribution is characterized by Poisson *marginals*.

Furthermore,

$$\text{Cov}(y_{ij}, y_{kl}) = \begin{cases} 0 & \text{for } i \neq k \\ \lambda_{ij} + \gamma & \text{for } i = k,\ j = l \\ \gamma & \text{for } i = k,\ j \neq l \end{cases}$$

The last equality follows since for $j \neq l$

$$\begin{aligned} \text{Cov}(z_{ij} + u_i, z_{il} + u_i) &= \text{Var}(u_i) \\ &= \gamma \end{aligned}$$

Normalization by the standard deviations of the two random variables yields the correlation form

$$\text{Corr}(y_{ij}, y_{kl}) = \frac{\gamma}{\sqrt{(\lambda_{ij} + \gamma)(\lambda_{kl} + \gamma)}} \tag{7.2}$$

It can be shown that the correlation coefficient cannot exceed the square root of the ratio of the smaller to the larger of the means of the two marginal distributions. The covariance matrix of y_i can be written in more compact form. Let $\Lambda_i = \text{diag}(\lambda_{ij})$ and $\mathbf{1}$ a $(J \times 1)$ vector of ones. The covariance matrix of $y_i = (y_{i1} \cdots y_{iJ})'$ is then given by

$$\text{Var}(y_i) = \Lambda_i + \gamma \mathbf{1}\mathbf{1}'$$

Since $\gamma \geq 0$ the model allows only non-negative correlations. This property of the MVP, as deplorable as it might look at first glance, is in fact a direct result of its simple one-factor structure. The same result would be obtained if z_{ij} and u_i were independently normally distributed with variances σ_z^2 and σ_u^2, respectively, from where $\text{Cov}(y_{ij}, y_{kl}) = \sigma_u^2 \geq 0$. But such random effects models are widely used in econometrics, in particular for panel data. Thus, the MVP model should (at least) be useful in related situations involving panel count data.

The joint probability function of the MVP distribution for individual i is given by

$$\begin{aligned} f(y_{i1}, ..., y_{iJ}) &= \exp\left[-(\lambda_{i1} + \ldots + \lambda_{iJ} + \gamma)\right] \\ &\quad \times \sum_{u_i=0}^{s_i} \frac{\gamma^{u_i}}{u_i!} \frac{\lambda_{i1}^{y_{i1}-u_i}}{(y_{i1}-u_i)!} \cdots \frac{\lambda_{iJ}^{y_{iJ}-u_i}}{(y_{iJ}-u_i)!} \end{aligned} \tag{7.3}$$

with $s_i = \min(y_{i1}, ..., y_{iJ})$. The intuition behind this joint probability function is as follows. First, u_i cannot exceed any of the y_{ij}'s because each count is the sum of u_i and a non-negative count z_{ij}. Hence, its upper bound is s_i. Secondly, the joint probability of observing $(y_{i1}, ..., y_{iJ})$ is the sum over the joint probabilities $f(u_i, y_{i1} - u_i, ..., y_{iJ} - u_i)$ where $u_i = 0, \ldots, s_i$. From independence, it follows that the joint probability can be factored such that

$$f(u_i, y_{i1} - u_i, ..., y_{iJ} - u_i) = f(u_i) f(y_{i1} - u_i) \cdots f(y_{iJ} - u_i)$$

In order to derive the conditional distributions of the MVP, consider for simplicity the bivariate case:

$$\begin{aligned} f(y_{i1}|y_{i2}) &= \frac{f(y_{i1}, y_{i2})}{f(y_{i2})} \\ &= \frac{\exp(-\lambda_{i1} - \lambda_{i2} - \gamma) \sum_{u_i=0}^{s_i} \frac{\gamma^{u_i}}{u_i!} \frac{\lambda_{i1}^{y_{i1}-u_i}}{(y_{i1}-u_i)!} \frac{\lambda_{i2}^{y_{i2}-u_i}}{(y_{i2}-u_i)!}}{\exp(-\lambda_{i2} - \gamma)(\lambda_{i2} + \gamma)^{y_{i2}}/y_{i2}!} \\ &= \sum_{u_i=0}^{s_i} \left\{ \frac{y_{i2}!}{u_i!(y_{i2} - u_i)!} \left(\frac{\gamma}{\lambda_{i2} + \gamma} \right)^{u_i} \left(\frac{\lambda_{i2}}{\lambda_{i2} + \gamma} \right)^{y_{i2}-u_i} \right. \\ &\qquad \left. \times \exp(-\lambda_{i1}) \frac{\lambda_{i1}^{y_{i1}-u_i}}{(y_{i1} - u_i)!} \right\} \end{aligned} \tag{7.4}$$

This is the distribution of the sum of two independent variables (Recall the generic formula for a convolution: $f(z) = \sum_{i=0}^{z} f_x(z - i) f_y(i)$). Here, y_{i1} is Poisson distributed with parameter λ_{i1}, and $u_i|y_{i2}$ is binomial distributed with $n = y_{i2}$ and $p = \gamma/(\lambda_{i2} + \gamma)$. It follows that

$$\begin{aligned} \mathrm{E}(y_{i1}|y_{i2}) &= \mathrm{E}(y_{i1}) + \mathrm{E}(u_i|y_{i2}) \\ &= \lambda_{i1} + \frac{\gamma}{\lambda_{i2} + \gamma} y_{i2} \end{aligned}$$

Thus, the bivariate Poisson distribution defines a linear regression between y_{i1} and y_{i2} (and conversely). This property could be used in order to test for correlated counts using OLS. If λ_{ij}, $j = 1, 2$ was specified as a non-linear function of additional parameters (such as $\lambda_{ij} = \exp(x'_{ij}\beta)$) the conditional expectation function would need to be estimated by non-linear least squares.

Interestingly, a slight modification of the conditional expectation function leads to a model that allows for both positive and negative correlations between y_{i1} and y_{i2} (and is thus unrelated to the BVP). Berkhout and Plug (2004) study the situation where y_{i1} is Poisson distributed and $y_{i2}|y_{i1}$ is also Poisson distributed with mean

$$\begin{aligned} \mathrm{E}(y_{i2}|y_{i1}) &= \lambda_{i2} \exp(\alpha y_{i1}) \\ &= \exp(x'_{i2}\beta_2 + \alpha y_{i1}) \end{aligned}$$

Thus, the conditioning variable enters multiplicatively rather than additively. They show that the sign of the correlation between y_{i1} and y_{i2} corresponds to the sign of α.

Probability Generating Function of the MVP

Joint distributions for non-negative integer random variables can be modeled using joint probability generating functions (See Appendix A). This method

provides an alternative characterization of the multivariate Poisson distribution. For notational convenience the exposition is limited to the bivariate case. The bivariate probability generating function of two random variable X and Y is defined as $\mathcal{P}(s_1, s_2) = \mathrm{E}(s_1^X s_2^Y)$. Thus, in the bivariate Poisson model, the probability generating function for the joint distribution of (y_{i1}, y_{i2}) is given by:

$$\begin{aligned}
\mathcal{P}(s_1, s_2) &= \mathrm{E}(s_1^{y_{i1}} s_2^{y_{i2}}) \\
&= \mathrm{E}(s_1^{z_{i1}+u_i} s_2^{z_{i2}+u_i}) \\
&= \mathrm{E}((s_1 s_2)^{u_i} s_1^{z_{i1}} s_2^{z_{i2}}) \\
&\overset{(*)}{=} \mathrm{E}((s_1 s_2)^{u_i}) \mathrm{E}(s_1^{z_{i1}}) \mathrm{E}(s_2^{z_{i2}}) \\
&= \exp(-\gamma + s_1 s_2 \gamma) \exp(-\lambda_{i1} + \lambda_{i1} s_1) \exp(-\lambda_{i2} + \lambda_{i2} s_2) \\
&= \exp(-\lambda_{i1} - \lambda_{i2} - \gamma + \lambda_{i1} s_1 + \lambda_{i2} s_2 + \gamma s_1 s_2)
\end{aligned} \tag{7.5}$$

where $(*)$ follows from the independence assumption.

The probability function can be derived from (7.5) using the relationship

$$f(y_{i1}, y_{i2}) = (y_{i1}! y_{i2}!)^{-1} \left. \frac{\partial^{y_{i1}+y_{i2}} \mathcal{P}}{(\partial s_1)^{y_{i1}} (\partial s_2)^{y_{i2}}} \right|_{s_1=s_2=0}$$

One can verify that this representation leads to the probability function (7.3). The marginal distributions are defined as (see Appendix A):

$$\begin{aligned}
\mathcal{P}^{(y_{i1})}(s_1) &= \mathcal{P}(s_1, 1) \\
&= \exp[-\lambda_{i1} - \gamma + (\lambda_{i1} + \gamma) s_1]
\end{aligned}$$

$$\begin{aligned}
\mathcal{P}^{(y_{i2})}(s_2) &= \mathcal{P}(1, s_2) \\
&= \exp[-\lambda_{i2} - \gamma + (\lambda_{i2} + \gamma) s_2]
\end{aligned}$$

The covariance of y_{i1} and y_{i2} can be calculated as

$$\begin{aligned}
\mathrm{Cov}(y_{i1}, y_{i2}) &= \mathrm{E}(y_{i1} y_{i2}) - \mathrm{E}(y_{i1}) \mathrm{E}(y_{i2}) \\
&= \gamma
\end{aligned}$$

since

$$\begin{aligned}
\mathrm{E}(y_{i1} y_{i2}) &= \sum_{y_{i1}=0}^{\infty} \sum_{y_{i2}=0}^{\infty} y_{i1} y_{i2} f(y_{i1}, y_{i2}) \\
&= \left[\frac{\partial^2 \mathcal{P}(s_1, s_2)}{\partial s_1 \partial s_2} \right]_{s_1=s_2=1} \\
&= \gamma + (\lambda_{i1} + \gamma)(\lambda_{i2} + \gamma)
\end{aligned}$$

For $\gamma = 0$, the probability generating function can be factored:

$$\mathcal{P}(s_1, s_2) = \mathcal{P}(s_1, 1) \mathcal{P}(1, s_2)$$

and, therefore, y_{i1} and y_{i2} are independent (See Appendix A). Like for the bivariate normal distribution, independence and no correlation are equivalent

notions. There are, however, two important differences: Whereas for the bivariate normal distribution both marginal *and* conditional distributions are again normal, here this holds only for the marginal distributions. Moreover, sums of Poisson random variables are again Poisson distributed if and only if they are independent: the probability generating function of the sum $y_{i1} + y_{i2}$ is obtained by setting $s_1 = s_2 = s$:

$$\mathcal{P}(s) = \exp[(\lambda_{i1} + \lambda_{i2})(s-1) + \gamma(s^2 - 1)]$$

For $\gamma = 0$, i.e. if the two Poisson variables are independent, this is the probability generating function of a Poisson distribution.

Bivariate Poisson Process

Yet another characterization of the bivariate Poisson distribution is based on the bivariate Poisson process: Let $y_1(t, t+\Delta), y_2(t, t+\Delta)$ be the number of events of two different types that occured between t and $t+\Delta$, $t, \Delta \in \mathbb{R}^+$. Assume that the probabilities of events y_1 or y_2 occuring in the interval $(t, t+\Delta)$ are independent of the previous process, and that

(i) The probability of one occurence of type 1 and no occurence of type 2 in the interval $(t, t+\Delta)$ is given by:

$$P(y_1 = 1, y_2 = 0) = \lambda_1 \Delta + \circ(\Delta)$$

(ii) The probability of one occurence of type 2 and no occurence of type 1 in the interval $(t, t+\Delta)$ is given by:

$$P(y_1 = 0, y_2 = 1) = \lambda_2 \Delta + \circ(\Delta)$$

(iii) The probability of one occurence of type 1 and one occurence of type 2 in the interval $(t, t+\Delta)$ is given by:

$$P(y_1 = 1, y_2 = 1) = \gamma \Delta + \circ(\Delta)$$

(iv) The probability of no event occuring is given by:

$$P(y_1 = 0, y_2 = 0) = 1 - \lambda_1 \Delta - \lambda_2 \Delta - \gamma \Delta + \circ(\Delta)$$

It can then be shown that the resulting probability generating function must be of the form

$$\mathcal{P}(t, s_1, s_2) = \exp[(-\lambda_1 - \lambda_2 - \gamma + \lambda_1 s_1 + \lambda_2 s_2 + \gamma s_1 s_2)t] \; . \tag{7.6}$$

Setting $t = 1$, the probability generating function for the bivariate Poisson distribution derived in (7.5) is obtained. This derivation of the bivariate Poisson distribution was proposed as early as 1926 by McKendrick (See the historical remarks in Kocherlakota and Kocherlakota, 1992). It can be given a spatial interpretation of moving along a Cartesian grid, where one-step movements along the y_1-axis and y_2-axis occur with probabilities λ_1 and λ_2, respectively, while a movement in both directions has probability γ.

Seemingly Unrelated Poisson Regression

The MVP probability model (7.3) together with parameterization

$$\lambda_{ij} = \exp(x'_{ij}\beta_j) - \gamma$$

is often refered to as seemingly unrelated Poisson regression. The model was introduced by King (1989a) who suggested estimation by maximum likelihood. Jung and Winkelmann (1993) give the first and second derivatives of the log-likelihood function.

Applications in econometrics include Jung and Winkelmann (1993) who study the joint determination of the number of voluntary and involuntary job changes over a ten-year period, and Ozuna and Gomez (1994) who study the number of trips to two recreational sites. Applications so far have been limited to the bivariate case, although this is definitely not a binding constraint.

Also, despite the labelling, all previous applications have dealt with data that are multivariate in nature rather than seemingly unrelated proper in the sense of Zellner (1962). This orientation has re-inforced the criticism of the MVP model as being potentially inappropriate, since it imposes non-negative correlation. This a-priori restriction is more of an drawback for multivariate data then it would be for SURE or panel data. In response, attention has shifted to multivariate mixing models such as the Poisson-log-normal model discussed below (Chib and Winkelmann, 2001, Gurmu and Elder, 1998).

Another criticism has been based on the restrictive variance assumption of the MVP model: the conditional expectation and conditional variance are assumed to be equal. One response has been to ignore the issue of over- or underdispersion in estimation but allow for valid inference by computing robust standard errors (Jung and Winkelmann, 1993). Alternatively, Winkelmann (2000a) derives a multivariate negative binomial model along the lines of the MVP model. This model allows for overdispersion. It is presented in the next chapter.

A final point of contention, raised by Gurmu and Elder (1998) is whether it is meaningful to assume that $z_{ij} \sim \text{Poisson}(\exp(x'_{ij}\beta_j) - \gamma)$. This specification does not guarantee that the parameter of the z_{ij}-distribution is positive, causing both conceptual and potentially numerical problems. In an alternative parameterization, $z_{ij} \sim \text{Poisson}(\exp(x'_{ij}\beta_j))$, resulting in a marginal distribution of $y_{ij} \sim \text{Poisson}(\exp(x'_{ij}\beta_j)+\gamma)$. Although the two models differ not only in their constant but also in the underlying assumption for the scedastic (variance) function, the interpretation of the regression parameters is the same in both parameterizations, as in either case $\partial \text{E}(y_{ij}|x_{ij})/\partial x_{ij} = \exp(x'_{ij}\beta_j)\beta_j$.

7.1.2 Multivariate Negative Binomial Model

A multivariate negative binomial (MVNB) model can be derived in close analogy to the MVP. Following Winkelmann (2000a), begin with a convolution structure and let

$$y_{ij} = z_{ij} + u_i$$

where z_{ij} and u_i have independent negative binomial distributions. In order to establish the distribution of the sum of two independent negative binomial distributions, recall the probability generating function of the negative binomial distribution from Chap. 2.3.1:

$$\mathcal{P}(s) = [1 + \theta(1 - s)]^{-\alpha}$$

In this specification, $\mathrm{E}(y) = \alpha\theta$ and $\mathrm{Var}(y) = \mathrm{E}(y)(1 + \theta)$. Thus, the sum of two independent negative binomial distributions is again negative binomial only if the two distributions share the common parameter θ. (This property of the negative binomial distribution was also exploited by Hausman, Hall and Griliches (1984, Appendix A), albeit in a different context). Consider a parameterization where

$$z_{ij} \sim \text{Negbin}(\theta = \sigma, \alpha = \lambda_{ij}/\sigma) \tag{7.7}$$

$$u_i \sim \text{Negbin}(\theta = \sigma, \alpha = \gamma/\sigma) \tag{7.8}$$

It follows that z_{ij} has mean $\lambda_{ij} = \exp(x'_{ij}\beta_j)$ and variance $\lambda_{ij}(1+\sigma)$, whereas u_i has mean γ and variance $\gamma(1 + \sigma)$. Thus, z_{ij} and u_i each are Negbin I distributed.

Applying the basic convolution rules to independent random variables, the distribution of $y_{ij} = z_{ij} + u_i$ can be established as

$$\begin{aligned} \mathcal{P}_y(s) &= \mathcal{P}_z(s)\mathcal{P}_u(s) \\ &= [1 + \sigma(1 - s)]^{-\lambda_{ij}/\sigma}[1 + \sigma(1 - s)]^{-\gamma/\sigma} \\ &= [1 + \sigma(1 - s)]^{-(\lambda_{ij}+\gamma)/\sigma} \end{aligned} \tag{7.9}$$

But (7.9) is the probability generating function of a Negbin I distribution with expectation $\mathrm{E}(y_{ij}) = \lambda_{ij} + \gamma$ and variance $\mathrm{Var}(y_{ij}) = (\lambda_{ij} + \gamma)(1 + \sigma)$. It is easy to verify that among the class of negative binomial distributions, only the Negbin I distribution is closed under convolution. The Negbin II distribution, in particular, is not.

Due to the common factor u_i, this model induces correlation between observations for the same individual but different outcomes: For $i = k$ and $j \neq l$

$$\begin{aligned} \mathrm{Cov}(y_{ij}, y_{kl}) &= \mathrm{Var}(u_i) \\ &= \gamma(1 + \sigma) \end{aligned}$$

Following the notation that was introduced for the MVP model, the covariance matrix of the MVNB model can be written in compact form as

$$\mathrm{Var}(y_i) = [\Lambda_i + \gamma\mathbf{1}\mathbf{1}'](1 + \sigma)$$

Note that this covariance matrix differs from the covariance matrix of the MVP model only by a factor of $(1 + \sigma)$. Thus, the MVNB model allows for overdispersion relative to the MVP model as long as $\sigma > 0$. The restriction implied by the MVP model ($\sigma = 0$) can be subject to test.

The joint probability function of the MVNB model for cluster i is obtained along the lines of (7.3):

$$f(y_{i1},\ldots,y_{iJ}) = \sum_{k=0}^{s_i} f_{\mathrm{NB}}(k) \prod_{j=1}^{J} f_{\mathrm{NB}}(y_{ij}-k) \tag{7.10}$$

where $s_i = \min(y_{i1},\ldots,y_{iJ})$ and f_{NB} is the Negbin I probability function. For instance, for $z_{ij} = y_{ij} - k$

$$f_{\mathrm{NB}}(z_{ij}) = \frac{\Gamma(\lambda_{ij}/\sigma + z_{ij})}{\Gamma(\lambda_{ij}/\sigma)\Gamma(z_{ij}+1)} \left(\frac{1}{1+\sigma}\right)^{\lambda_{ij}/\sigma} \left(\frac{\sigma}{1+\sigma}\right)^{z_{ij}} \tag{7.11}$$

The parameters of the model can be estimated by maximizing the corresponding log-likelihood function.

7.1.3 Multivariate Poisson-Gamma Mixture Model

An alternative approach to induce correlation amongst the counts has been pursued by Hausman, Hall and Griliches (1984) (see also Dey and Chung, 1992). In their model, correlation is generated by an individual specific multiplicative error term. The error term represents individual specific unobserved heterogeneity. The mixture multivariate density of $y_i = (y_{i1}\cdots y_{iJ})'$ is obtained after integration

$$f(y_i|x_i) = \int \left[\prod_{j=1}^{J} \frac{\exp(-\lambda_{ij}u_i)(\lambda_{ij}u_i)^{y_{ij}}}{\Gamma(y_{ij}+1)}\right] g(u_i)\mathrm{d}u_i \tag{7.12}$$

If u_i is gamma distributed with $\mathrm{E}(u_i) = 1$ and $\mathrm{Var}(u_i) = \alpha^{-1}$ it can be shown that the joint distribution function of y_i is of a negative binomial form with distribution function.

$$\begin{aligned} f(y_i|x_i) &= \left(\prod_{j=1}^{J} \frac{(\lambda_{ij})^{y_{ij}}}{\Gamma(y_{ij}+1)}\right) \frac{\alpha^\alpha}{\Gamma(\alpha)} \int e^{-u_i(\lambda_{i.}+\alpha)} u^{y_{i.}+\alpha-1}\mathrm{d}u_i \\ &= \left(\prod_{j=1}^{J} \frac{(\lambda_{ij})^{y_{ij}}}{\Gamma(y_{ij}+1)}\right) \frac{\Gamma(y_{i.}+\alpha)}{\Gamma(\alpha)} \alpha^\alpha (\lambda_{i.}+\alpha)^{-(y_{i.}+\alpha)} \end{aligned} \tag{7.13}$$

where $y_{i.} = \sum_{j=1}^{J} y_{ij}$ and $\lambda_{i.} = \sum_{j=1}^{J} \lambda_{ij}$. Note that this model is very closely related to the univariate Poisson-gamma mixture leading to the univariate negative binomial distribution. The only difference is that mixing is over a common variable u_i rather than over independent gamma variable u_{ij}. The similarity is also seen in the marginals of the multivariate Poisson-gamma model that are univariate negative binomial with $\mathrm{E}(y_{ij}) = \lambda_{ij}$ and $\mathrm{Var}(y_{ij}) = \lambda_{ij} + \gamma\lambda_{ij}^2$ where $\gamma = \alpha^{-1}$ (i.e., of the Negbin II variety).

The covariance between outcomes for a given individual can be derived as follows:

$$\begin{aligned}\mathrm{Cov}(y_{ij}, y_{il}) &= \mathrm{E}_u\mathrm{Cov}(y_{ij}, y_{il}|u_i) + \mathrm{Cov}_u[\mathrm{E}(y_{ij}|u_i), \mathrm{E}(y_{il}|u_i)|u_i] \\ &= 0 + \mathrm{Cov}_u(\lambda_{ij}u_i, \lambda_{il}u_i) \\ &= \gamma\lambda_{ij}\lambda_{il}, \quad j \neq l\end{aligned}$$

In compact form, the covariance matrix for individual i is given by

$$\mathrm{Var}(y_i) = \Lambda_i + \Lambda_i\gamma\mathbf{1}\mathbf{1}'\Lambda_i$$

where $\Lambda_i = \mathrm{diag}(\lambda_{ij})$ as before.

Hence, the multivariate Poisson-gamma model allows for overdispersion, and within-individual correlation. As for the MVP and MVNB models, the covariances are non-negative. In contrast to the two previous models, the multivariate Poisson-gamma model does not have an "equi-covariance" property. Rather, within individual covariances are an increasing function of the product of the expected values λ_{ij} and λ_{il}. This could be a useful feature for modeling non-negative random variables. In particular, it eliminates the strict upper bound to the correlation that was observed for the MVP distribution.

A potential disadvantage of this model is that the covariances are not determined independently of the dispersion. Hence, a finding of a significant γ can be as much an indicator of overdispersion in the data as it might be an indicator of correlation (or both). In the MVNB model, by contrast, these two features of the data can be identified, and thus estimated, separately. We also note that all multivariate models discussed so far require covariances to be non-negative. Depending on the application, this can be an undesirable feature, and a more general model is discussed in the next chapter.

Finally, note that for $J = 2$ the multivariate Poisson-gamma mixture model of Hausman, Hall and Griliches (1984) is identical to the bivariate negative binomial model attributed to Marshall and Olkin (1990) (See also Munkin and Trivedi, 1999). Its joint probability distribution function is given by

$$\begin{aligned}f(y_1, y_2|x_1, x_2) = \frac{\Gamma(y_1 + y_2 + \alpha)}{y_1!y_2!\Gamma(\alpha)} &\left(\frac{\lambda_1}{\lambda_1 + \lambda_2 + 1}\right)^{y_1} \left(\frac{\lambda_2}{\lambda_1 + \lambda_2 + 1}\right)^{y_2} \\ &\times \left(\frac{1}{\lambda_1 + \lambda_2 + 1}\right)^{\alpha}\end{aligned}$$

which is easily seen to be a special case of (7.13).

7.1.4 Multivariate Poisson-Log-Normal Model

Assume that conditionally on a $(J \times 1)$ vector of individual and outcome specific random effects $\varepsilon_i = (\varepsilon_{i1} \cdots \varepsilon_{iJ})'$ the distribution of y_i is independent Poisson

$$f(y_i|\varepsilon_i) = \prod_{j=1}^{J} f_p(y_{ij}|\lambda_{ij}\exp(\varepsilon_{ij})) \tag{7.14}$$

where f_p is the Poisson distribution function. Since ε_i is unobserved, the model is not complete. In analogy to the univariate Poisson-log-normal model, Aitchison and Ho (1989) suggested a multivariate extention where ε_i is J-variate normal distributed

$$f(\varepsilon_i|\Omega) = \phi_J(-0.5\text{diag}\Omega, \Omega) \tag{7.15}$$

and Ω is the covariance matrix. Aside from the random effects, the model is thus characterized by the parameters $\psi = (\lambda, \Omega)$. The importance of the non-zero mean specification depending on the diagonal elements of Ω is explained below. Aitchison and Ho (1989), as well as Good and Pirog-Good (1989) who considered a bivariate Poisson-log-normal distribution, restricted their attention to the case without regressors but the model can be readily extended to the case with regressors by letting $\lambda_{ij} = \exp(x'_{ij}\beta)$.

To understand this specificationn note that the conditional mean and variance of the outcomes are given by

$$\text{E}(y_{ij}|\varepsilon_{ij}) = \text{Var}(y_{ij}|\varepsilon_{ij}) = \lambda_{ij}\exp(\varepsilon_{ij})\,.$$

This allows one to derive the expectation and variance of the marginal joint distribution of y_i without integration. A simple reparameterization facilitates the analysis. Let $u_{ij} = \exp(\varepsilon_{ij})$ and $u_i = (u_{i1} \cdots u_{iJ})'$. The assumption on ε_i implies that $u_i \sim \text{MVLN}_J(\mathbf{1}, \Sigma)$, a multivariate log-normal distribution with mean vector $\mathbf{1}$ and covariance matrix Σ where $\sigma_{ij} = \exp(\omega_{ij}) - 1$ and thus $\Sigma = \exp(\Omega) - \mathbf{1}\mathbf{1}'$. Hence,

$$y_{ij}|\lambda_{ij}, u_{ij} \sim \text{Poisson}(\lambda_{ij}u_{ij})$$

and the model is in the form of a Poisson-log-normal distribution.

To derive the marginal moments, let $\lambda_i = (\lambda_{i1} \cdots \lambda_{iJ})'$ and $\Lambda_i = \text{diag}(\lambda_i)$. Then by the law of the iterative expectations one obtains

$$\text{E}(y_i|\lambda_i, \Omega) = \lambda_i \tag{7.16}$$

and

$$\text{Var}(y_i|\Lambda_i, \Omega) = \Lambda_i + \Lambda_i[\exp(\Omega) - \mathbf{1}\mathbf{1}']\Lambda_i \tag{7.17}$$

Hence, the covariances between the counts are represented by the terms

$$\text{Cov}(y_{ij}, y_{kl}) = \lambda_{ij}(\exp(\omega_{jl}) - 1)\lambda_{kl},\ j \neq l,\ i = k$$

which can be positive or negative depending on the sign of ω_{jl}, the (j, l) element of Ω. The correlation structure of the counts is thus unrestricted. Moreover, the model allows for overdispersion as long as $\omega_{ii} > 0$. Note, however, that the marginal distribution of the counts y_i cannot be obtained by direct computation, requiring as it does the evaluation of a J -variate integral of the Poisson distribution with respect to the distribution of ε_i

$$f(y_i|\lambda_i, \Omega) = \int \prod_{j=1}^{J} f_p(y_{ij}|\lambda_{ij}, \varepsilon_{ij})\phi(\varepsilon_i| - 0.5\text{diag}\Omega, \Omega)\text{d}\varepsilon_i \tag{7.18}$$

where f as above is the Poisson probability function conditioned on $(\lambda_{ij}, \varepsilon_{ij})$ and ϕ is the J-variate normal distribution. This J-dimensional integral cannot be solved in closed form for arbitrary Ω. For $J = 2$, Munkin and Trivedi (1999) discuss estimation by simulated maximum likelihood. However, this is strictly speaking not necessary as one could obtain a one-dimensional integral through a factorization of ϕ into a conditional and a marginal distribution and then apply Gauss-Hermite quadrature. A simulation method based on Markov chain Monte Carlo that works well for high-dimensional problems, is presented in Chap. 7.1.4.

If Ω is a diagonal matrix, the J-variate integral reduces to the product of J single integrals

$$f(y_i|\lambda_i, \Omega) = \prod_{j=1}^{J} \int f_p(y_{ij}|\lambda_{ij}, \varepsilon_{ij})\phi(\varepsilon_{ij}| - 0.5\omega_{jj}, \omega_{jj})\text{d}\varepsilon_{ij} \tag{7.19}$$

a product of J independent univariate Poisson-log normal densities. See Chap. 4.2.3 for a discussion of the univariate Poisson-log-normal model. For $\Omega = 0$ the joint probability simplifies to a product of J independent Poisson densities.

Discussion

There are several ways to generalize the multivariate Poisson log-normal model. The considerations are similar to those for selectivity models. First, the marginal distribution of ε_{ij} may be known but not normal. In this case, one can apply results in Lee (1983) and Weiss (1999) to generate a multivariate distribution in which the random variables are allowed to correlate (see also Chap. 5.2.2). If $F(\varepsilon_{ij})$ is the cumulative marginal distribution function of ε_{ij}, then the transformed random variable

$$v_{ij} = \Phi^{-1}(F(\varepsilon_{ij})),$$

where Φ^{-1} is the inverse cumulative density function of the standard normal distribution, is standard normal distributed. To introduce correlation, assume that the joint distribution $f(v_i) = \phi_J(0, D)$ is multivariate normal with covariance matrix D. Clearly, ε_i is multivariate normal only if F is the normal distribution. In particular, D is usually not the covariance matrix of ε_i. As such, the interpretation of the covariance structure is somewhat difficult. However, this set-up has generated a multivariate distribution for ε_i with known marginal cumulative distribution functions equal to $F(\cdot)$ and unrestricted covariance structure. No application of this method to multivariate count data is known at this stage.

An alternative possibility is to relax the strong distributional assumptions. A method based on squared polynomial series expansions for the unknown

density of the correlated errors is suggested in Gurmu and Elder (1998). The method was originally developed by Gurmu, Rilstone and Stern (1998) for the univariate case. Gurmu and Elder extend the method to the bivariate case. There is some doubt whether this method could be succesfully applied to high dimensional multivariate data.

A final alternative is to abandon distributional assumption altogether and specify first and second order moments of the joint distribution of $u_i = \exp(\varepsilon_i)$, and thus y_i, only. This is discussed in Chap. 7.1.6.

7.1.5 Latent Poisson-Normal Model

A latent Poisson-normal model for bivariate correlated counts is presented in van Ophem (1999). In this model, count data are interpreted as realizations of an underlying (latent) normally distributed variable. One problem is that the support of count data distributions is unbounded. To make it a well defined problem, assume that $y = 0, 1, \ldots, K$, where K is an upper bound. This restriction does not matter for estimation, however, as actual data are always finite (van Ophem, 1999).

Consider the following mapping from a standard normal variable u_1 to the count variable y_1:

$$y_1 = k \text{ iff } \eta_{k-1} < u_1 < \eta_k \quad k = 0, 1, \ldots, K$$

where $\eta_{-1} = -\infty$ and $\eta_K = \infty$. Thus

$$P(y_1 = k) = \Phi(\eta_k) - \Phi(\eta_{k-1})$$

and

$$P(y_1 \leq k) = \Phi(\eta_k)$$

or, conversely,

$$\eta_k = \Phi^{-1}[P(y_1 \leq k)]$$

This relationship defines η_k uniquely for any marginal distribution $P(y_1 = k)$. So far, the model has only been re-parameterized without changing its substance. However, now assume that for a second count variable y_2, a similar procedure gives

$$\mu_m = \Phi^{-1}[P(y_2 \leq m)]$$

If u_1 and u_2 are bivariate normal with correlation ρ, then we can write the joint cumulative probability function as

$$P(y_1 \leq k, y_2 \leq m) \int_{-\infty}^{\eta_k} \int_{-\infty}^{\mu_m} \phi(u_1, y_2; \rho) \mathrm{d}u_2 \mathrm{d}u_1$$

where ϕ is the bivariate normal density with means 0, variances 1 and correlation ρ. Moreover, the likelihood contribution can be calculated as

$$\begin{aligned} P(y_1 = k, y_2 = m) = {} & P(y_1 \leq k, y_2 \leq m) - P(y_1 \leq k-1, y_2 \leq m) \\ & -P(y_1 \leq k, y_2 \leq m-1) \\ & +P(y_1 \leq k-1, y_2 \leq m-1) \end{aligned} \tag{7.20}$$

To summarize, this model has well-specified marginal distributions, in this case Poisson, and a correlation structure that allows both for positive and negative correlations. The correlation results from a latent bivariate normal distribution with correlation ρ. If the only goal of the analysis is to allow for negative correlations it is not clear whether this model offers an advantage over the relatively simpler Poisson-log-normal model. The Poisson-log-normal model has no Poisson marginal distributions, though. Hence the latent Poisson-normal model has an advantage if one strongly believes in Poisson marginals. In practice, however, these restrictive marginals are likely to speak against rather than for this model.

A serious limitation of the latent Poisson-normal model is its focus on the bivariate case. Extentions to higher dimensional multivariate data appear impractical. A major advantage of the model is its versatility: the approach can easily be adopted to any bivariate discrete random variable.

7.1.6 Moment-Based Methods

A parametric model with correlated errors was introduced in Chap. 7.1.4, where it was assumed that

$$y_{ij}|\lambda_{ij}, u_{ij} \sim \text{Poisson}(\lambda_{ij} u_{ij})$$

and

$$u_i = \begin{pmatrix} u_{i1} \\ \vdots \\ u_{iJ} \end{pmatrix} \sim \text{MVLN}(\mathbf{1}, \Sigma)$$

where MVLN denotes the multivariate normal distribution with expected value $\mathbf{1}$ and covariance matrix Σ. This model is appropriate for panel data, where Σ could reflect either serial or contemporaneous correlations, as well as for genuine multivariate data. Maximum likelihood estimation of this model in general requires simulation methods.

Alternatively, Gourieroux, Monfort and Trognon (1984b) discuss semiparametric estimation of this model. Their approach differs slightly from the one employed here, as it is based on a bivariate Poisson distribution with common additive factor (i.e., in their model there are two sources of intra-cluster correlation, one being the common additive factor and the other being mixing over correlated errors). However, this is an inconsequential complication that can be dropped for ease of exposition. For the same reason, we follow Gourieroux, Monfort and Trognon (1984b) and focus on the bivariate case. Hence, the model can be written as

$$\begin{pmatrix} y_{i1}|u_{i1} \\ y_{i2}|u_{i2} \end{pmatrix} \sim \text{ independently Poisson } \begin{pmatrix} \exp(x'_{i1}\beta_1)u_{i1} \\ \exp(x'_{i2}\beta_2)u_{i2} \end{pmatrix}$$

with

$$\mathrm{E}(u_{i1}) = \mathrm{E}(u_{i2}) = 1$$

and

$$\mathrm{Var}\begin{pmatrix} u_{i1} \\ u_{i2} \end{pmatrix} = \Sigma$$

Hence, only the first two moments of the mixing distribution are specified. Let $\lambda_{i1} = \exp(x'_{i1}\beta_1)$ and $\lambda_{i2} = \exp(x'_{i2}\beta_2)$. The correlated random effects introduce within cluster correlation among $y_i = (y_{i1}, y_{i2})'$ as

$$\mathrm{Var}\begin{pmatrix} y_{i1} \\ y_{i2} \end{pmatrix} = \begin{pmatrix} \lambda_{i1} & 0 \\ 0 & \lambda_{i2} \end{pmatrix} + \begin{pmatrix} \lambda_{i1} & 0 \\ 0 & \lambda_{i2} \end{pmatrix} \Sigma \begin{pmatrix} \lambda_{i1} & 0 \\ 0 & \lambda_{i2} \end{pmatrix} \tag{7.21}$$

is not a diagonal matrix. Gourieroux, Monfort and Trognon (1984b) suggest estimating β_1 and β_2 by non-linear least squares minimizing

$$\sum_{i=1}^{n} \left[(y_{i1} - \exp(x'_{i1}\beta_1))^2 + (y_{i2} - \exp(x'_{i2}\beta_2))^2 \right]$$

The estimators $\hat{\beta}_1$ and $\hat{\beta}_2$ can be used to define the residuals

$$\hat{w}_{i1} = y_{i1} - \exp(x'_{i1}\hat{\beta}_1)$$

$$\hat{w}_{i2} = y_{i2} - \exp(x'_{i2}\hat{\beta}_2)$$

Moreover, consistent estimators of σ_{11}, σ_{12} and σ_{22}, the elements of Σ, are obtained by applying ordinary least squares to

$$\hat{w}_{i1}^2 - \exp(x'_{i1}\hat{\beta}_1) = \sigma_{11}\exp(2x'_{i1}\hat{\beta}_1) + \text{ disturbance}$$

$$\hat{w}_{i2}^2 - \exp(x'_{i2}\hat{\beta}_2) = \sigma_{22}\exp(2x'_{i2}\hat{\beta}_2) + \text{ disturbance}$$

$$\hat{w}_{i1}\hat{w}_{i2} = \sigma_{12}\exp(x'_{i1}\hat{\beta}_1)\exp(x'_{i2}\hat{\beta}_2) + \text{ disturbance}$$

Having obtained an estimator $\hat{\Sigma}$, Gourieroux, Monfort and Trognon (1984b) recommend the use of quasi-generalized pseudo maximum likelihood (QGPML) in order to reach the lower bound of the asymptotic covariance matrix of pseudo maximum likelihood estimators. QGPML solves the weighted non-linear least squares problem

$$\min \sum_{i=1}^{n} (y_{i1} - \lambda_{i1}, y_{i2} - \lambda_{i2}) \left[\widehat{\mathrm{Var}} \begin{pmatrix} y_{i1} \\ y_{i2} \end{pmatrix} \right]^{-1} \begin{pmatrix} y_{i1} - \lambda_{i1} \\ y_{i2} - \lambda_{i2} \end{pmatrix}$$

where

$$\widehat{\mathrm{Var}} \begin{pmatrix} y_{i1} \\ y_{i2} \end{pmatrix}$$

is obtained from (7.21) using $\hat{\sigma}$, $\hat{\beta}_1$ and $\hat{\beta}_2$.

7.1.7 Copula Functions

In statistics, a copula is a multivariate joint distribution function defined on the n-dimensional unit cube $[0, 1]$ such that every marginal distribution is uniform on the interval $[0, 1]$. For example, the Gaussian copula, for $n = 2$, is

$$P(U \leq u, V \leq v) = C(u, v) = \Phi_2(\Phi^{-1}(u), \Phi^{-1}(v); \rho)$$

Two other examples are Clayton's copula

$$C(u, v) = (u^{-\theta} + v^{-\theta} - 1)^{-1/\theta}$$

and the Frank copula

$$C(u, v) = -\theta^{-1} \log \left\{ 1 + \frac{(e^{-\theta u} - 1)(e^{-\theta u} - 1)}{(e^{-\theta} - 1)} \right\}$$

The marginal distributions implied by bivariate copulas are

$$F(u) = P(U \leq u, V \leq 1) = C(u, 1)$$

and

$$F(v) = P(U \leq 1, V \leq v) = C(1, v)$$

respectively. It is easy to verify that all three copulas have the key property that their marginal distributions are uniform, as $C(u, 1) = u$ and $C(1, v) = v$.

The significance of copulas lies in the fact that by way of transformation, any joint distribution function can be expressed as a copula applied to the marginal distributions. This result is due to Sklar. Sklar's theorem states that given a joint distribution function $F(y_1, \ldots, y_k)$, and respective marginal distribution functions, there exists a copula C such that the copula binds the margins to give the joint distribution.

For the bivariate case, Sklar's theorem can be stated as follows. For any bivariate distribution function $F(y_1, y_2)$, let $F_1(y_1) = F(y_1, \infty)$ and $F_2(y_2) = F(\infty, y_2)$ be the univariate marginal probability distribution functions. Then there exists a copula C such that

$$F(y_1, y_2) = C(F_1(y_1), F_2(y_2))$$

Moreover, if the marginal distributions are continuous, the copula function C is unique. We see, that the copula is now expressed as a function of distribution functions (cdf's). But a standard result in statistics states that cdf's are uniform distributed over the interval $[0, 1]$. Since the marginal distributions of a copula are uniform distributed, it follows that the marginal distribution of $y_1 = F_1^{-1}(u)$ and $y_2 = F_2^{-1}(v)$ are F_1 and F_2, as stated.

The practical significance of copula functions in empirical modeling stems from the fact that they can be used to build new multivariate models for given univariate marginal component cdf's. If the bivariate cdf $F(y_1, y_2)$ is unknown, but the univariate marginal cdf's are of known form, then one can choose a copula function and thereby generate a representation of the

unknown joint distribution function. The key is that this copula function introduces dependence, captured by additional parameter(s), between the two random variables (unless the independence copula $C(u, v) = uv$ is chosen). The degree and type of dependence depends on the choice of copula. There is a large literature on this topic (Trivedi and Zimmer, 2007).

It is known that the copula representation is not unique in the case of discrete random variables. Zimmer and Trivedi (2006) express the view that this non-uniqueness is not a problem in practice for using copulas in count data modeling. Otherwise, one might also follow Denuit and Lambert (2005) and transform count data into continuous responses by adding a realization from a standard uniform $u[0, 1]$ distribution to each count y, and apply then continuous data copula methods for estimation and inference.

Without "continued" count data, the joint probability function for two count data with marginal distributions $F_1(y_1)$, $F_2(y_2)$ and copula function C is given by

$$\begin{aligned} P(y_1, y_2) = {} & C(F_1(y_1), F_2(y_2); \theta) - C(F_1(y_1 - 1), F_2(y_2); \theta) \\ & - C(F_1(y_1), F_2(y_2 - 1); \theta) + C(F_1(y_1 - 1), F_2(y_2 - 1); \theta) \end{aligned}$$

The marginal distributions can be made functions of regressors X. For example, if we consider the case of two Poisson marginals,

$$F_j(y_j) = \sum_{k=0}^{y_j} \frac{e^{-\lambda_j} \lambda_j^k}{k!}$$

where

$$\lambda_j = \exp(x' \beta_j).$$

Such a model, combined with a Gaussian copula, is discussed by van Ophem (1999). Van Ophem (2000) uses the Gaussian copula to generate a count data model with endogenous binary regressor. Zimmer and Trivedi (2006) model simultaneity between insurance choice and two measures (counts) of health care utilization using the trivariate Frank copula.

7.2 Panel Data Models

Panel data provide repeated measurements (over time) on dependent and independent variables for a sample of individuals or households or firms. Examples are pre- and post treatment studies in a controlled experiment, such as the number of epileptic seizures before and after treatment (Diggle, Liang, and Zeger, 1995). In social sciences, panel count data, such as the number of days absent from work in a given year, for a number of years, are observed in repeated linked household surveys, for example the U.S. Panel Study of Income Dynamics or the German Socio Economic Panel. So far, the leading application of panel count data models in the econometrics literature is to

firm level data on patent numbers (Hausman, Hall and Griliches, 1984). Recent contributions include Cincera (1997), Crépon and Duguet (1997a, 1997b), Blundell, Griffith and van Reenen (1995, 1999), and Montalvo (1997). Other examples include the number of doctor visits (Geil et al., 1997, Winkelmann, 2004a) and the number of workdays lost in a panel of U.S. manufacturing establishments (Ruser, 1991).

Methods for panel count data differ from standard count data models in at least one of three ways. First, they address the non-standard form of the covariance matrix of the observations that arises since the assumption of independent observations is most likely invalid. Second, and relatedly, they provide a richer framework for addressing the issue of unobserved heterogeneity than do univariate count data (see Chap. 4). In particular, dependence between the unobserved heterogeneity and the regressors is no longer excluded. Third, models for panel count data allow the introduction of dynamic elements, such as a lagged dependent variable, into the regression part of the model.

Panel methods typically also differ from genuine multivariate count data models. First, panel data models are usually somewhat more restrictive in their covariance structure, as they frequently assume that dependence is generated by unobserved heterogeneity that is specific to the individual but constant over time. Secondly, panel data models explicitly consider the possibility that the unobserved individual heterogeneity factor is correlated with one or more explanatory variables. In this situation, conditional models are required.

To illustrate the type of modeling issues encountered for panel data, consider the determinants of patent numbers. It is likely that differences in technological opportunities or operating skills may affect the observed number of patents. And yet, these firm specific factors are typically not captured by explanatory variables. If firm specific unobservables are correlated over time, a plausible assumption to start with, they will cause a positive correlation among the repeated observations of a single firm. One special, and most commonly assumed, case is that of a time-invariant firm effect. This can be seen as a limiting case of correlated effects, where the correlation is perfect. In addition, such a firm effect may be correlated with explanatory variables. By construction, this must be so in a dynamic context, where a lagged dependent variable is included among the regressors. But correlation, i.e. endogeneity, can arise in other situations as well.

There are three basic approaches for dealing with individual specific effects in panel count data:

1. Robust methods and pseudo maximum likelihood
2. Parametric random effects models
3. Fixed effects models

As in the linear model, the first two methods work whenever the individual effects are independent of the regressors (the assumption of absence of correlation is in general not sufficient in the context of non-linear count data models). The issue is then one of correct inference versus efficient estimation.

An example for an inefficient but robust method is the Poisson pseudo likelihood estimator (see Chap. 3.3.3) which retains consistency despite a non-standard covariance structure implied by the individual error, as long as the conditional expectation is correctly specified. An example for an efficient random effects model for count data is the panel negative binomial model to be introduced later in this chapter. It is very similar to the standard negative binomial model, i.e., it arises from a Poisson distribution with gamma heterogeneity. The difference is that the mixing over the gamma distribution is now not done observation wise, but rather jointly for the block of all observations for one individual over all time periods, as discussed in Chapter 7.1.3.

Both methods fail if the unobserved individual specific effect is correlated with explanatory variables. To see why this is so, consider the following conditional expectation function

$$\mathrm{E}(y_{it}|x_{it}, \alpha_i) = \exp(x'_{it}\beta)\alpha_i \tag{7.22}$$

where $i = 1, \ldots, N$ indexes the individual (or household, or firm), $t = 1, \ldots, T$ indexes time, and α_i is a time invariant individual specific error term. The conditional expectation $\mathrm{E}(y_{it}|x_{it})$ is therefore given by

$$\mathrm{E}(y_{it}|x_{it}) = \exp(x'_{it}\beta)\mathrm{E}(\alpha_i|x_{it})$$

where $\mathrm{E}(\alpha_i|x_{it}) \neq \textit{constant}$ if x is endogenous. In this case, pooled Poisson or random effects estimation is inappropriate, and we should rather use a model that conditions on α_i and estimates (7.22) directly. If we embed the conditional expectation function (7.22) in a Poisson probability model, we obtain the fixed effects Poisson model.

7.2.1 Fixed Effects Poisson Model

Let $\lambda_{it} = \exp(x'_{it}\beta)$. Then the Poisson model with multiplicative individual specific fixed effect has conditional probability function

$$f(y_{it}|x_{it}, \alpha_i) = \frac{\exp(-\alpha_i\lambda_{it})(\alpha_i\lambda_{it})^{y_{it}}}{y_{it}!} \tag{7.23}$$

The fixed effects model treats the α_i's as parameters that need to be estimated jointly with β. The advantages of the fixed effects model are twofold:

1. the population distribution of α_i does not need to be specified. This avoids inconsistency of a misspecied random effects model;
2. the individual specific error term α_i may be correlated with the explanatory variables x_{it}.

In order to estimate a fixed effects Poisson model one could simply include n individual specific dummy variables, that is, one intercept for each individual. This may be hard or impossible when N is large and T is small as is the case in many applications: $(N+k)$ parameters need to be estimated. For large N this is likely to exceed software restrictions. However, an inspection of the

log-likelihood function reveals that this problem does not arise in the fixed effects Poisson model as analytical expressions for α_i can be derived and used to concentrate the likelihood as a function of β (see Cameron and Trivedi, 1998, and Blundell, Griffith, and Windmeijer, 2002).

We start with the assumption that the regressors x_{it} are strictly exogenous, such that the T observations for one individual are independent conditional on α_i, and we can write

$$f(y_{i1},\ldots,y_{iT}|x_{i1},\ldots,x_{iT},\alpha_i) = f(y_{i1}|x_{i1},\alpha_i)\cdots f(y_{iT}|x_{iT},\alpha_i)$$

If we define vectors $y_i = (y_{i1},\ldots,y_{iT})'$ and $x_i = (x_{i1},\ldots,x_{iT})'$, we obtain, using (7.23)

$$\begin{aligned} f(y_i|\alpha_i,x_i) &= \prod_{t=1}^{T} \exp(-\alpha_i\lambda_{it})(\alpha_i\lambda_{it})^{y_{it}}/y_{it}! \\ &= \exp\left(-\alpha_i\sum_{t=1}^{T}\lambda_{it}\right)\prod_{t=1}^{T}\alpha_i^{y_{it}}\prod_{t=1}^{T}\lambda_{it}^{y_{it}}/\prod_{t=1}^{T}y_{it}! \end{aligned} \tag{7.24}$$

The log-likelihood contribution of individual i is therefore

$$\ell_i(\alpha_i,\beta) = -\alpha_i\sum_{t=1}^{T}\lambda_{it} + \ln\alpha_i\sum_{t=1}^{T}y_{it} + \sum_{t=1}^{T}y_{it}\ln\lambda_{it} - \sum_{t=1}^{T}\ln y_{it}! \tag{7.25}$$

with first derivative

$$\frac{\partial\ell_i(\alpha_i,\beta)}{\partial\alpha_i} = -\sum_{t=1}^{T}\lambda_{it} + \alpha_i^{-1}\sum_{t=1}^{T}y_{it}$$

Therefore, the maximum likelihood estimator for α_i is

$$\hat{\alpha}_i = \frac{\sum_{t=1}^{T}y_{it}}{\sum_{t=1}^{T}\lambda_{it}} = \frac{\bar{y}_i}{\bar{\lambda}_i} \tag{7.26}$$

This solution makes intuitive sense. Whereas in the linear model with additive fixed effect, the fixed effects are estimated by the difference of $\bar{y}_i$ and $\bar{\hat{y}}_i$, in the exponential model with multiplicative fixed effect, they are estimated by the corresponding ratio.

We can now substitute this expression back into (7.25) in order to obtain the *concentrated* log-likelihood function (i.e. the likelihood function that depends no longer on α_i). Taking into account all N observations, this is

$$\begin{aligned} \ell^c(\beta) &= \sum_{i=1}^{N}\left\{-\sum_{t=1}^{T}y_{it} + \left(\ln\frac{\sum_{t=1}^{T}y_{it}}{\sum_{t=1}^{T}\lambda_{it}}\right)\sum_{t=1}^{T}y_{it} + \sum_{t=1}^{T}y_{it}\ln\lambda_{it} - \sum_{t=1}^{T}\ln y_{it}!\right\} \\ &= constant + \sum_{i=1}^{N}\left\{\sum_{t=1}^{T}y_{it}\ln\lambda_{it} - \sum_{t=1}^{T}y_{it}\ln\sum_{t=1}^{T}\lambda_{it}\right\} \end{aligned} \tag{7.27}$$

where the *constant* collects all terms that do not depend on β. We can now take derivatives with respect to β.

$$\frac{\partial \ell^c(\beta)}{\partial \beta} = \sum_{i=1}^{N} \left\{ \sum_{t=1}^{T} y_{it} x_{it} - \frac{\sum_{t=1}^{T} y_{it}}{\sum_{t=1}^{T} \lambda_{it}} \sum_{t=1}^{T} \lambda_{it} x_{it} \right\}$$
$$= \sum_{i=1}^{N} \sum_{t=1}^{T} x_{it} \left(y_{it} - \frac{\sum_{t=1}^{T} y_{it}}{\sum_{t=1}^{T} \lambda_{it}} \lambda_{it} \right)$$

such that the first-order condition for β can thus be written independently of α_i as

$$\sum_{i=1}^{N} \sum_{t=1}^{T} \left(y_{it} - \frac{\bar{y}_i}{\bar{\lambda}_i} \lambda_{it} \right) x_{it} \equiv 0 \tag{7.28}$$

The maximum likelihood estimator for β is the value $\hat{\beta}$ that solves (7.28). This result has a number of noteworthy properties. First, there is no "incidental parameter problem" in the Poisson model with multiplicative fixed effects. In other words, the parameters of the conditional expectation function, β, can be estimated consistently for fixed T, as long as $N \to \infty$. This aspect distinguishes the fixed effects Poisson model from the fixed effects logit model, say, where such a problem arises.

Second, if we compare the first-order condition of the fixed effects Poisson model to the first order condition to the pooled Poisson maximum likelihood estimator,

$$\sum_{i=1}^{N} \sum_{t=1}^{T} (y_{it} - \lambda_{it}) x_{it} \equiv 0$$

we see that both of them establish a zero-correlation condition between residuals and the regressors. However, the fixed effects Poisson model uses *scaled* residuals whereas the pooled estimator uses unscaled residuals $y_{it} - \lambda_{it}$.

Third, we find that consistent estimation of β does not require that the dependent variable is truly Poisson distributed. Rather, a simple moment condition, namely that

$$\mathrm{E}(y_{it} | x_{i1}, \ldots, x_{iT}, \alpha_i) = \alpha_i \lambda_{it}$$

is sufficient for consistent estimation of β. As in the simple cross-section model, therefore, the Poisson model has a pseudo-likelihood interpretation. This approach is discussed further in Chapter 7.2.2. Before that, it will be shown that an identical first-order condition is obtained from a conditional likelihood approach.

Conditional Maximum Likelihood

Hausman, Hall and Griliches (1984) suggested to estimate the fixed effects Poisson model by conditioning the likelihood contribution of individual i on the individual specific sum $\sum_{t=1}^{T} y_{it}$. Since observations are independently

Poisson distributed conditional on α_i, the distribution of this sum is Poisson distributed as well

$$\sum_{t=1}^{T} y_{it} \sim \text{Poisson}\left(\alpha_i \sum_{t=1}^{T} \lambda_{it}\right)$$

Next, consider the joint density for the i-th individual. We can write

$$f(y_{i1}, \ldots, y_{iT}) = f\left(y_{i1}, \ldots, y_{iT}, \sum_{t=1}^{T} y_{it}\right)$$

since the sum of counts for individual i over time is fully determined by its components and thus adds no new information. Thus, for the conditional distribution

$$\begin{aligned} f\left(y_{i1}, \ldots, y_{iT} \middle| \sum_{t=1}^{T} y_{it}\right) &= \frac{f(y_{i1}, \ldots, y_{iT}, \sum_{t=1}^{T} y_{it})}{f(\sum_{t=1}^{T} y_{it})} \\ &= \frac{e^{-\sum_t \lambda_{it}} \prod_t \lambda_{it}^{y_{it}}}{\prod_t y_{it}!} \Big/ \frac{e^{-\sum_t \lambda_{it}} (\sum_t \lambda_{it})^{\sum_t y_{it}}}{(\sum_t y_{it})!} \end{aligned} \tag{7.29}$$

Terms involving α_i have canceled out. The resulting probability expression turns out to be of a multinomial form, with conditional probabilities proportional to

$$f\left(y_{i1}, \ldots, y_{iT} \middle| \sum_{t} y_{it}\right) \propto \prod_{t=1}^{T} \left(\frac{\lambda_{it}}{\sum_{t=1}^{T} \lambda_{it}}\right)^{y_{it}}$$

Upon taking logarithms, we find that the log likelihood function of this model is exactly the same as the concentrated log likelihood (7.27). Thus, first-order conditions and the maximum likelihood estimator are identical as well. There is no difference between the two approaches.

7.2.2 Moment-based Estimation of the Fixed Effects Model

In the previous Chapter, the full conditional distribution of $f(y_{it}|x_{it}, \alpha_i)$ was specified, in this case the Poisson distribution. However, this assumption was unnecessarily strong, as β can be estimated based on a simple moment restriction as well, namely

$$\mathrm{E}(y_{it}|x_{i1}, \ldots, x_{iT}, \alpha_i) = \alpha_i \lambda_{it} \tag{7.30}$$

This condition, defining strict exogeneity of x, implies that

$$\mathrm{E}(y_{it}|x_{i1}, \ldots, x_{iT}) = \lambda_{it} \mathrm{E}(\alpha_i|x_{i1}, \ldots, x_{iT})$$

and

$$\mathrm{E}(\bar{y}_i|x_{i1}, \ldots, x_{iT}) = \bar{\lambda}_i \mathrm{E}(\alpha_i|x_{i1}, \ldots, x_{iT})$$

Therefore,

$$\mathrm{E}\left(y_{it} - \frac{\bar{y}_i}{\bar{\lambda}_i}\lambda_{it} \middle| x_{i1}, \ldots, x_{iT} \right)$$
$$= \lambda_{it}\mathrm{E}(\alpha_i | x_{i1}, \ldots, x_{iT}) - \lambda_{it}\mathrm{E}(\alpha_i | x_{i1}, \ldots, x_{iT}) = 0$$

which in turn implies that

$$\mathrm{E}\left\{ \left(y_{it} - \frac{\bar{y}_i}{\bar{\lambda}_i}\lambda_{it} \right) x_{it} \right\} = 0 \tag{7.31}$$

The moment estimator for β that solves (7.28), the orthogonality condition for the mean scaled residuals, is therefore consistent. Blundell, Griffith and Windmeijer (2002) call this model a "within groups mean scaling model". In this situation, moment estimator, Poisson maximum likelihood estimator and Poisson conditional likelihood estimator are all the same.

Note that it is not enough to assume $\mathrm{E}(y_{it}|x_{it}, \alpha_i) = \alpha_i\lambda_{it}$, because this does not imply $\mathrm{E}(\bar{y}_i|x_{it}, \alpha_i) = \alpha_i\bar{\lambda}_i$, since

$$\mathrm{E}(\bar{y}_i|x_{it}, \alpha_i) = T^{-1}\sum_{s=1}^{T}\mathrm{E}(y_{is}|x_{it}, \alpha_i)$$

involves $T-1$ terms with $s \neq t$ that are not specified unless strict exogeneity is assumed. Also, it is true that the strict exogeneity condition 7.30 implies $T-1$ additional moment restrictions, in addition to (7.31), such as

$$\mathrm{E}\left\{ \left(y_{it} - \frac{\bar{y}_i}{\bar{\lambda}_i}\lambda_{it} \right) x_{is} \right\} = 0 \qquad s \neq t$$

If these conditions were actually used, one would leave the Poisson PML framework and could proceed by GMM estimation of this overidentified model. However, such an approach appears to have not been implemented so far in the literature.

Within the pseudo likelihood framework, the estimated standard errors need to be adjusted accordingly in order to obtain valid inference. The covariance matrix can be estimated consistently using

$$\mathrm{Var}_{\mathrm{PML}}(\hat{\beta}) = \hat{\mathcal{J}}^{-1}\hat{\mathcal{I}}\hat{\mathcal{J}}^{-1}$$

where

$$\hat{\mathcal{J}} = \sum_{i=1}^{n}\left(\sum_{t=1}^{T} x_{it}x_{it}'\frac{\bar{y}_i}{\bar{\lambda}_i}\lambda_{it} - \frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T} x_{it}x_{is}'\frac{\bar{y}_i}{\bar{\lambda}_i}\lambda_{it}\lambda_{is} \right)$$

and

$$\hat{\mathcal{I}} = \sum_{i=1}^{n}\sum_{t=1}^{T}\sum_{s=1}^{T} x_{it}x_{is}'\left(y_{it} - \frac{\bar{y}_i}{\bar{\lambda}_i}\lambda_{it} \right)\left(y_{is} - \frac{\bar{y}_i}{\bar{\lambda}_i}\lambda_{is} \right)'$$

(See Cameron and Trivedi, 1998).

Overall, this robustness property is a very useful aspect of the Poisson fixed effects model. Thus one does not need to worry about overdispersion, or

other expressions of "non-Poissoness". This is even more of an advantage here, in the panel case, relative to the cross-section case, as extensions of the fixed effects approach to other parametric models, such as the negative binomial, while in some special case technically feasible, suffer nevertheless from some serious limitations, as we will see in the next Chapter.

It should also be pointed out that none of the results of this Chapter carry over to generalizations of the Poisson model that lead to modified conditional expectation functions, such as truncated Poisson models, hurdle or with-zero models, and the like. The development of panel data models and fixed effects estimators for such generalized models is still an open task.

7.2.3 Fixed Effects Negative Binomial Model

In the presence of overdispersion, a potentially more efficient estimator can be based on the fully parametric fixed effects negative binomial model that was introduced by Hausman, Hall and Griliches (1984). They discuss estimation of the model by conditional maximum likelihood. As for the fixed effects Poisson model, the conditioning is on the individual specific sums $\sum_{t=1}^{T} y_{it}$. In order to derive a closed form expression for the joint conditional probability distribution for individual i, it is necessary that the probability distribution of $\sum_{t=1}^{T} y_{it}$ can be expressed in closed form. As shown in Chap. 7.1.2, a sum of independent negative binomial random variables is again negative binomial distributed if and only if the component distributions are of Negbin I type with probability generating function

$$\mathcal{P}(s) = [1 + \theta(1 - s)]^{-\delta}$$

and common parameter θ. Constrained by this requirement, Hausman, Hall and Griliches (1984) suggest the parameterization $\delta = \lambda_{it}$ and $\theta = \theta_i$, an individual specific fixed effect. In this parameterization, the probability function for observation y_{it} is given by

$$f(y_{it}) = \frac{\Gamma(\lambda_{it} + y_{it})}{\Gamma(\lambda_{it})\Gamma(y_{it} + 1)} \left(\frac{1}{1 + \theta_i}\right)^{\lambda_{it}} \left(\frac{\theta_i}{1 + \theta_i}\right)^{\lambda_{it}} \tag{7.32}$$

To fully appreciate the role of the individual specific effect θ_i in this model, note that

$$\mathrm{E}(y_{it}|\theta_i) = \lambda_{it}\theta_i$$

and

$$\mathrm{Var}(y_{it}|\theta_i) = \lambda_{it}(\theta_i + \theta_i^2) = \mathrm{E}(y_{it}|\theta_i)(1 + \theta_i)$$

Thus, this Negbin I-type model introduces a time invariant variance-to-mean ratio. With $\theta_i = \exp(\alpha_i)$, we could as well write

$$\mathrm{E}(y_{it}|\alpha_i) = \exp(\alpha_i + x_{it}'\beta)$$

and

$$\mathrm{Var}(y_{it}|\alpha_i) = \exp(\alpha_i + x'_{it}\beta)(1 + \exp(\alpha_i))$$

Hence, the α'_is are not just differential intercepts in the mean function – they also appear also as a separate shifter in the variance function. From this, it follows that the α'_is play a different role than x_{it}, and it becomes logically impossible to interpret these terms as a representation of omitted explanatory variables. This aspect limits the usefulness of the model for use in genuine panel count data applications.

In order to preserve the standard structure of a fixed effects panel data model, one might be tempted to let instead $\delta = \exp(\alpha_i + x'_{it}\beta)$. Unfortunately, this parameterization is unsuitable for computational reasons as α_i fails then to drop out of the conditional likelihood function. But this contravenes the purpose of the whole exercise.

Putting aside these caveats regarding the interpretation of the model for a moment, we will now show that the θ'_is indeed disappear from the individual specific likelihood contribution. First, for a given individual i, the y_{it} are independent over time, such that

$$f\left(\sum_{t=1}^{T} y_{it}\right) = \frac{\Gamma(\sum_t \lambda_{it} + \sum_t y_{it})}{\Gamma(\sum_t \lambda_{it})\Gamma(\sum_t y_{it} + 1)} \left(\frac{1}{1+\theta_i}\right)^{\sum_t \lambda_{it}} \left(\frac{\theta_i}{1+\theta_i}\right)^{\sum_t y_{it}}$$

while

$$f(y_{i1}, \ldots, y_{iT}) = \prod_{t=1}^{T} \frac{\Gamma(\lambda_{it} + y_{it})}{\Gamma(\lambda_{it})\Gamma(y_{it} + 1)} \left(\frac{1}{1+\theta_i}\right)^{\lambda_{it}} \left(\frac{\theta_i}{1+\theta_i}\right)^{y_{it}}$$

Therefore,

$$f\left(y_{i1}, \ldots, y_{iT} \left| \sum_{t=1}^{T} y_{it}\right.\right) = \frac{\Gamma(\sum_t \lambda_{it})\Gamma(\sum_t y_{it} + 1)}{\Gamma(\sum_t \lambda_{it} + \sum_t y_{it})} \times \prod_{t=1}^{T} \frac{\Gamma(\lambda_{it} + y_{it})}{\Gamma(\lambda_{it})\Gamma(y_{it} + 1)} \tag{7.33}$$

All terms involving θ_i have cancelled out of the conditional model, and the remaining parameters can be easily estimated. This model is available in the current releases of econometric software packages such as Stata (StataCorp., 1999) and Limdep (Greene, 1995).

7.2.4 Random Effects Count Data Models

Random effects model should be used if the assumption of independence between the individual specific effects and the regressors appears tenable. In this case, a random effect model will tend to be more efficient. Relative to the fixed effects model, it has N additional degrees of freedom. Moreover, random effects model have the advantage that time invariant regressors can

be included. The independence assumptions can be tested using Hausman's (1978) method (see also (3.90).

Hausman, Hall and Griliches (1984) discuss the Poisson model with gamma distributed individual specific effect $u_i = \exp(\varepsilon_i)$. The derivation of this model is very similar to that of the univariate negative binomial model in Chap. 4.3. The difference is that unobserved heterogeneity is now individual specific, that is, modeled as u_i rather than u_{it}. As shown in Chap. 7.1.3,if u_i is independently gamma distributed with parameters (γ, γ) (i.e., with mean 1 and variance $1/\gamma$) the joint marginal distribution of $y_i = (y_{i1}, \ldots, y_{iT})'$ is of negative binomial form with

$$f(y_i) = \frac{\Gamma(\sum_t y_{it} + \gamma)}{\Gamma(\gamma)} \left(\frac{\gamma}{\gamma + \sum_t \lambda_{it}}\right)^{\gamma} \frac{1}{(\gamma + \sum_t \lambda_{it})^{\sum_t y_{it}}} \prod_{t=1}^{T} \left(\frac{\lambda_{it}^{y_{it}}}{y_{it}!}\right)$$

In addition, Hausman, Hall and Griliches (1984) have introduced a random effects negative binomial model. As for the fixed effects negative binomial model, the starting point is a Negbin I model as in (7.32). Now, assume that $1/(1+\theta_i)$ is distributed as beta(a, b). With this assumption, θ_i can be integrated out and, after some algebra, the resulting joint probability function for individual i can be written as

$$f(y_i) = \frac{\Gamma(a+b)\Gamma(a+\sum_t \lambda_{it}))\Gamma(b+\sum_t y_{it})}{\Gamma(a)\Gamma(b)\Gamma(a+b+\sum_t \lambda_{it}) + \sum_t y_{it})} \times \prod_t \frac{\Gamma(\lambda_{it}+y_{it})}{\Gamma(\lambda_{it})y_{it}!} \tag{7.34}$$

Moment-Based Methods

Brännäs and Johansson (1996) consider moment based estimation of a panel data model with both serially *and* contemporaneously correlated errors. Let

$$\mathrm{E}(y_{it}|\lambda_{it}, u_{it}) = \mathrm{Var}(y_{it}|\lambda_{it}, u_{it}) = \lambda_{it} u_{it}$$

Denote by Σ_{ii} the variance matrix of $u_i = (u_{i1} \cdots u_{iT})'$ where i indexes individuals and t indexes time. Hence, Σ_{ii} contains the within-individual serial correlations. Some restrictions, such as covariance stationarity of a $AR(1)$ process, can be imposed. Similarly, Σ_{ij} is the covariance matrix between u_i and u_j, $i \neq j$. The diagonal elements of Σ_{ij} are contemporaneous correlations. Thus, the classical SURE model is obtained if both Σ_{ii} and Σ_{ij} are diagonal matrices. For $\Sigma_{ij} = 0$, this model is a multivariate extension of Zeger's (1988) time-series model (see Chap. 7.3).

Under the assumptions of the general model,

$$\mathrm{Var}(y_i) = \Lambda_i + \Lambda_i \Sigma_{ii} \Lambda_i$$

where $\Lambda_i = \mathrm{diag}(\lambda_{it})$ as before. However, in addition,

$$\mathrm{Cov}(y_i, y_j) = \Lambda_i \Sigma_{ij} \Lambda_j \qquad i \neq j$$

Brännäs and Johansson (1996) estimate the parameters of the model by GMM.

7.2.5 Dynamic Panel Count Data Models

There has been substantial recent interest in methods for panel count data with correlated individual specific effects and weakly exogenous regressors. The literature includes Montalvo (1997), Crépon and Duguet (1997a), Blundell, Griffith and van Reenen (1995) and Blundell, Griffith and Windmeijer (2002). With correlated individual specific effects, estimation requires the use of fixed effects. It was shown in Chap. 7.2.1 that the fixed effects Poisson estimator solving the first-order conditions

$$\sum_{i=1}^{N}\sum_{t=1}^{T} x_{it}\left(y_{it} - \frac{\lambda_{it}}{\overline{\lambda}_i}\bar{y}_i\right) = 0 \tag{7.35}$$

is consistent if all regressors are strictly exogenous. This excludes the presence of predetermined regressors, or weakly exogenous regressors, such as lagged dependent variables. Consider $x_{it} = y_{i,t-1}$. The conditional expectation

$$\mathrm{E}(y_{it}|x_{i1},\ldots,x_{iT},\alpha_i) = \mathrm{E}(y_{it}|y_{i0},\ldots,y_{i,T-1},\alpha_i)$$

becomes ill defined for $t < T$. For example, $\mathrm{E}(y_{it}|y_{it}) = y_{it}$, and this approach makes little sense. Instead, we need to consider a useful exogeneity definition for predetermined regressors, such as

$$\mathrm{E}(y_{it}|x_{i1},\ldots,x_{it},\alpha_i) = \alpha_i\lambda_{it}$$

The "problem" with this definition is that under weak exogeneity, the moment condition

$$\mathrm{E}\left\{\left(y_{it} - \frac{\bar{y}_i}{\overline{\lambda}_i}\lambda_{it}\right)x_{it}\right\} = 0$$

no longer holds, and the Poisson fixed effects estimator based on (7.35) is therefore not a consistent estimator. The reason is the same as the one mentioned in Chapter 7.2.2 regarding the insufficiency of the assumption that $\mathrm{E}(y_{it}|x_{it}) = \alpha_i\lambda_{it}$: weak exogeneity is not sufficient to determine

$$\mathrm{E}(\bar{y}_i|x_{i1},\ldots,x_{it},\alpha_i)$$

Thus, alternative methods are required. The problem is to find a transformation that eliminates the multiplicative fixed effect and at the same time generates useable moment conditions. Following Chamberlain (1992), Blundell, Griffith and van Reenen (1995) and Montalvo (1997), consider the alternative of scaling residuals by future observations, or leads. Define

$$v_{it} = y_{it} - \frac{y_{i,t+1}}{\lambda_{i,t+1}}\lambda_{it} \tag{7.36}$$

Here, we consider scaling by the first lead observation, although scaling by any y_{is} and λ_{is}, $s > t$ would be possible as well (see Montalvo, 1997, for a more general formulation). Under the weak exogeneity assumption, we have

$$\mathrm{E}(y_{it} - \alpha_i\lambda_{it}|x_{i1},\ldots,x_{it},\alpha_i) = 0$$

Similarly,

$$\begin{aligned}
&\mathrm{E}(y_{i,t+1} - \alpha_i \lambda_{i,t+1} | x_{i1}, \ldots, x_{it}, \alpha_i) \\
&= \mathrm{E}_{x_{i,t+1}}[\mathrm{E}(y_{i,t+1} - \alpha_i \lambda_{i,t+1} | x_{i1}, \ldots, x_{it}, x_{i,t+1}, \alpha_i) \\
&= \mathrm{E}_{x_{i,t+1}}[0] = 0
\end{aligned}$$

It follows that

$$\begin{aligned}
&\mathrm{E}\left(y_{it} - \frac{y_{i,t+1}}{\lambda_{i,t+1}} \lambda_{it} \middle| x_{i1}, \ldots, x_{it} \right) \\
&= \lambda_{it} \mathrm{E}(\alpha_i | x_{i1}, \ldots, x_{it}) - \frac{\lambda_{i,t+1} \mathrm{E}(\alpha_i | x_{i1}, \ldots, x_{it})}{\lambda_{i,t+1}} \lambda_{it} = 0
\end{aligned}$$

Thus, an estimator mimicking the moments conditions under strict exogeneity would solve

$$\sum_{i=1}^{N} \sum_{t=1}^{T} x_{it} \left(y_{it} - \frac{y_{i,t+1}}{\lambda_{i,t+1}} \lambda_{it} \right) = 0 \tag{7.37}$$

Alternatively, all prior values $x_{i,t-s}, s \geq 1$ can be used as instruments as well. This leads then to a GMM estimator as in Montalvo (1997): Define v_i to be the vector

$$v_i = \begin{bmatrix} y_{i1} - y_{i2} \exp[(x_{i1} - x_{i2})'\beta] \\ y_{i2} - y_{i3} \exp[(x_{i2} - x_{i3})'\beta] \\ \\ y_{iT-1} - y_{iT} \exp[(x_{iT-1} - x_{iT})'\beta] \end{bmatrix}$$

and a matrix of instruments Z_i as

$$Z_i = \begin{bmatrix} z_{i1} & 0 & \cdots & 0 \\ 0 & z_{i2} & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & z_{iT-1} \end{bmatrix}$$

where $z_{it} = (1, x_{i1}, \ldots, x_{it})$. Contrary to the case in which the variables are strictly exogenous, in this case there is no common set of instruments. Instead, the set increases with the number of periods. The GMM estimator of β is obtained by minimizing

$$\left[\sum_{i=1}^{n} v_i' Z_i \right] \widehat{W}_n^{-1} \left[\sum_{i=1}^{n} Z_i' v_i \right]$$

where the weight matrix is given by

$$\widehat{W}_n = \frac{1}{n} \sum_{i=1}^{n} Z_i' \hat{v}_i \hat{v}_i' Z_i$$

Alternatively, Wooldridge (1997a) proposes to eliminate the fixed effects using the transformation

$$u_{it}^{\dagger} = \frac{y_{it}}{\lambda_{it}} - \frac{y_{it+1}}{\lambda_{it+1}} \tag{7.38}$$

which equals v_{it} divided by λ_{it}. Windmeijer (2000) shows that this transformation is applicable for endogenous as well as weakly exogenous regressors. He also suggests that the failure of the Wooldridge transformation for non-negative regressors (in which case the associated β would go to infinity) can be overcome by transforming the x's first around their grand mean.

An alternative approach to deal with weakly endogenous regressors in panel count data is pursued by Blundell, Griffith and Windmeijer (2002) who use pre-sample information to form instruments for GMM estimation based on the mean-scaling model.

7.3 Time-Series Count Data Models

Pure time series count data can be seen as a special case of panel count data where $n = 1$ and T is large. Examples from the previous count data literature include the number of strikes per month (Buck, 1984), the number of bank failures per year (Davutyan, 1989) and the founding rate of national labor unions (Barron, 1992). In practice, the absence of a cross-sectional dimension makes a substantial difference, and developments of specialist time series models have been pursued independently of, and in most cases preceeding, those of panel models. The main concern of this literature has been a parsimonious and yet flexible correlation structure.

Dependence across time periods can be modeled in one of two ways. The first way is the introduction of an explicit lag structure in the endogenous count variable. This approach is also referred to as an "observation-driven" model (Firth 1991). The alternative is a "parameter-driven" model where time-series characteristics are introduced by correlated unobserved heterogeneity, following Zeger (1988), who augments the Poisson model by a multiplicative error term that follows an autoregressive process. This approach introduces both overdispersion and autocorrelation into y_t. Zeger proposes estimation of the model parameters by quasi-likelihood in the tradition of generalized linear models.

The observation driven approach is pursued by Al-Osh and Alzaid (1987, 1988) who define a fully parametric framework for modelling integer valued process with serial correlation. Al-Osh and Alzaid (1987) considers the case of integer valued autoregression, whereas Al-Osh and Alzaid (1988) deals with integer valued moving averages. In either case is the transition model characterized through a stopped-sum distribution (i.e., "binomial mixing" or "binomial thinning"). A synthesis of the two approaches that combines the INAR(1) structure with additional dependence from correlated errors is proposed by Brännäs and Hellström (2002).

An extensive survey of these methods, including an analysis of the performance of the estimators in simulation studies, is provided by Jung (1999). Ronning and Jung (1992), Brännäs (1994), and Böckenholt (1999) give applications of integer valued modeling in econometrics. See also Jung and Liesenfeld (2001).

Time Series with Correlated Multiplicative Errors

A time series model with correlated multiplicative error was proposed by Zeger (1988). This model can be seen as a special case of the multivariate Poisson model with correlated errors that was discussed in Chap. 7.1.4 and in Chap. 7.1.6. Recall that in the multivariate Poisson-log-normal model

$$\mathrm{E}(y_{it}|u_{it}) = \lambda_{it}u_{it}$$

and

$$\mathrm{Var}(y_{it}|u_{it}) = \lambda_{it}u_{it}$$

where

$$u_i = (u_{i1} \cdots u_{iT})' \sim \mathrm{MVLN}(\mathbf{1}, \Sigma)$$

For a time series, $n = 1$. Without further assumptions, the parameters of this model cannot be identified from a pure time-series. For instance, Σ, a symmetric $(T \times T)$ matrix, has $T(T+1)/2$ different elements. While an unrestricted covariance matrix can be estimated with multivariate data, restrictions are needed for time series data. For instance, Zeger (1988) considers a covariance stationary process where $\mathrm{Cov}(u_t, u_{t+\tau}) = \sigma_u(\tau)$. Under this assumption,

$$\Sigma = \begin{pmatrix} \sigma^2 & \sigma(1) & \dots & \sigma(t-1) \\ \sigma(1) & \sigma^2 & \dots & \sigma(t-2) \\ \vdots & \vdots & \ddots & \vdots \\ \sigma(t-1) & \sigma(t-2) & \dots & \sigma^2 \end{pmatrix} \tag{7.39}$$

Zeger (1988) adopts a quasi-, rather than maximum, likelihood framework. Thus, the assumption of log-normality is dropped, and only the first two moments of the distribution of y_{it} are taken into consideration. Zeger's approach is based on the score function

$$D'V^{-1}(y - \lambda) = 0 \tag{7.40}$$

where $D = \mathrm{d}\lambda/\mathrm{d}\beta$ is of dimension $(T \times k)$, y and λ are of dimension $(T \times 1)$, and

$$V = \mathrm{Var}(y) = \Lambda + \Lambda\Sigma\Lambda$$

where $\Lambda = \mathrm{diag}(\lambda_t)$. For independent observations, V is a diagonal matrix with diagonal element λ_t and the score equations reduce to the sum of the individual scores. In a time series context, however, Σ has non-zero off-diagonal elements as specified above.

The estimator that solves (7.40) has the well defined asymptotic distribution of a quasi-likelihood estimator under arbitrary forms of the covariance matrix (See, for instance McCullagh and Nelder, 1989, Chap. 9, Zeger, 1988). In practice, V is unknown and thus requires estimation. Zeger (1988) suggests the moment estimators

$$\hat{\sigma}^2 = \frac{\sum_{t=1}^{T} \tilde{y}_t^2 - \hat{\lambda}_t}{\sum_{t=1}^{T} \tilde{y}_t^2}$$

and

$$\hat{\sigma}(\tau) = \frac{\sum_{t=\tau+1}^{T} \tilde{y}_t \tilde{y}_{t-\tau}}{\hat{\sigma}^2 \sum_{t=\tau+1}^{T} \hat{\lambda}_t \hat{\lambda}_{t-\tau}}$$

where $\tilde{y}_t = y_t - \hat{\lambda}_t$. An iterative process can be used for estimation.

Alternatively, one may want to approximate the general covariance matrix Σ, and thus V, by a more parsimonious parameterization that follows for instance from a stationary autoregressive process. In this way, one can also avoid the repeated computation of the inverse of V, a matrix of dimension $(T \times T)$. First, note that V can be rewritten as

$$\begin{aligned} V &= (\Lambda + \sigma^2 \Lambda^2)^{1/2} R (\Lambda + \sigma^2 \Lambda^2)^{1/2} \\ &= D^{1/2} R D^{1/2} \end{aligned}$$

where R is the autocorrelation matrix of u and $D = \Lambda + \sigma^2 \Lambda^2$. Take, for instance, the case where u_t is assumed to follow a first-order autoregression (This case is also discussed in Wun, 1991). Then,

$$R^{-1} = \frac{1}{1-\rho} \begin{pmatrix} 1 & -\rho & 0 & \dots & 0 \\ -\rho & 1+\rho^2 & -\rho \dots & 0 & \\ 0 & -\rho & 1+\rho^2 & -\rho \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & -\rho & 1 \end{pmatrix}$$

(or, alternatively, $R^{-1} = L'L$ where L is the matrix that applies the autoregressive filter, i.e., $Ly = y_t + \alpha y_{t-1}$, $t > 1$). Thus, the inverse of V can be computed as

$$V^{-1} = D^{-1/2} R^{-1} D^{-1/2}$$

which is a considerable simplification.

Brännäs and Johansson (1994) consider estimation of the model by pseudo maximum likelihood. As long as the mean function is correctly specified, the Poisson model remains consistent but the asymptotic covariance matrix of the estimator needs to be adjusted.

Integer Valued Autoregression

Another model for time series count data is the integer valued autoregressive model (INAR), due to Al-Osh and Alzaid (1987) and McKenzie (1988) (see also Ronning and Jung, 1992). A random variable y follows a first order INAR process with Poisson marginals (written $y \sim$INAR(1)) if

$$y_t \stackrel{\mathrm{d}}{=} \alpha \circ y_{t-1} + \varepsilon_t \tag{7.41}$$

where

$$y_{t-1} \sim \text{Poisson}(\lambda)$$

$$\varepsilon_t \sim \text{Poisson}((1-\alpha)\lambda)$$

$$\varepsilon_t, y_{t-1} \text{ independent}$$

$$\alpha \circ y_{t-1} = \sum_{i=1}^{y_{t-1}} d_i$$

$$\alpha \in [0,1]$$

and

$$\{d_i\} \text{ i.i.d. with } P(d_i = 1) = 1 - P(d_i = 0) = \alpha\,.$$

The symbol "$\stackrel{\mathrm{d}}{=}$" stands for "is equally distributed as". Equation (7.41) defines a somewhat unusual relationship as y_t is a random variable even as α, y_{t-1}, and ε_t are known. In the remainder of this part, equality signs will have the same interpretation, although the explicit notation using "d" is dropped for simplicity.

In (7.41), $\alpha \circ y_{t-1}$ is a mixture of a binomial distribution and a Poisson distribution. For independent d_i and y_{t-1}, the mixture operation $'\circ'$ is called *binomial thinning* (McKenzie, 1988). It replaces the scalar multiplication in the continuous AR(1) model. $\alpha \circ y_{t-1}$ denotes the number of elements out of $t-1$ that survive to period t. The probability of survival is given by α. By the rules for convolutions (See Appendix A) $\alpha \circ y_{t-1} \sim \text{Poisson}(\alpha\lambda)$.

This model has the following interpretation: the innovation process $\{\varepsilon_t\}$ gives the number of new elements entering the process. The total number of elements in t is the sum of surviving and newly entering elements with marginal distribution $y_t \sim \text{Poisson}(\lambda)$. (The INAR(1) process has the following properties:

i) $0 \circ y = 0,\ 1 \circ y = y$
ii) $\mathrm{E}(\alpha \circ y) = \alpha \mathrm{E}(y)$
iii) $\underbrace{\alpha \circ \cdots \circ \alpha}_{k\text{-times}} \circ y = \alpha^k \circ y$

From (7.41) and ii), it follows that

$$\mathrm{E}(y_t | y_{t-1}) = \alpha y_{t-1} + (1-\alpha)\lambda\,. \tag{7.42}$$

Like for the first order autoregressive process with normally distributed innovations, the conditional expectation of y_t is linear in y_{t-1}. However, the regression is not linear in the parameters. Also, there is an additional source of randomness: given ε_t and y_{t-1}, y_t is still a (displaced binomial distributed) random variable.

Using iii) and recursive substitution, (7.41) can be rewritten as:

$$\begin{aligned} y_t &= \alpha \circ y_{t-1} + \varepsilon_t \\ &= \alpha \circ (\alpha \circ y_{t-2} + \varepsilon_{t-1}) + \varepsilon_t \\ &= \alpha \circ (\alpha \circ (\alpha \circ y_{t-3} + \varepsilon_{t-2}) + \varepsilon_{t-1}) + \varepsilon_t \\ &\quad \vdots \end{aligned}$$

i.e.,

$$y_t = \alpha^\tau \circ y_{t-\tau} + \sum_{j=0}^{\tau-1} \alpha^j \circ \varepsilon_{t-j} \,. \tag{7.43}$$

The marginal distribution of the INAR(1) process is then given by

$$y_t = \sum_{j=0}^{\infty} \alpha^j \circ \varepsilon_{t-j} \,. \tag{7.44}$$

The effect of $\{\varepsilon_t\}$ on y_t is reduced exponentially with increasing lag length. (7.43) implies for the auto-covariance structure:

$$\begin{aligned} \gamma(\tau) &= \mathrm{Cov}(y_{t-\tau}, y_t) \\ &= \mathrm{Cov}\left(y_{t-\tau}, \alpha^\tau \circ y_{t-\tau}\right) + \mathrm{Cov}\left(y_{t-k}, \sum_{j=0}^{\tau-1} \alpha^j \circ \varepsilon_{t-j}\right) \\ &\overset{(ii)}{=} \alpha^\tau \mathrm{Var}(y_{t-\tau}) + \sum_{j=0}^{\tau-1} \alpha^j \mathrm{Cov}(y_{t-\tau}, \varepsilon_{t-j}) \\ &= \alpha^\tau \gamma(0) \end{aligned}$$

The auto-correlations $\rho(\tau) = \gamma(\tau)/\gamma(0)$ are, in contrast to those of the Gaussian process, restricted to the positive interval (0,1). The INAR(1)-Poisson process is stationary for $\alpha \in (0,1)$. For $y_0 \sim \mathrm{Poisson}(\lambda)$ it holds $\forall t$ that

$$\begin{aligned} &\mathrm{E}(y_t) = \lambda \\ &\mathrm{Cov}(y_t, y_{t-\tau}) = \alpha^\tau \lambda \;, \tau = 0, 1, \ldots \end{aligned}$$

In particular, for $\tau = 0$, the typical Poisson property of equidispersion follows. Estimation can proceed by maximum likelihood. The INAR(1) model has Markovian property

$$f(y_t|y_{t-1}, y_{t-2}, \ldots) = f(y_t|y_{t-1})$$

and thus the joint distribution of the sample can be factored as

$$f(y_t, y_{t-1}, y_{t-2}, \ldots) = f(y_t|y_{t-1}) f(y_{t-1}|y_{t-2}) \ldots f(y_0)$$

The conditional distribution of y_t given y_{t-1} is a binomial-Poisson mixture, the probabilities of which are given by

$$f(y_t|y_{t-1}) = \exp[-(1-\alpha)\lambda](1-\alpha)^{y_{t-1}+y_t}\lambda^{y_t} \times \sum_{k=0}^{\min(y_t, y_{t-1})} \frac{\alpha^k y_{t-1}!}{(1-\alpha)^{2k}\lambda^k k!(y_t-k)!(y_{t-1}-k)!}$$

Denoting the factor in the second line by B_t, the joint distribution of the process can be written as

$$f(y_0, y_1, \ldots, y_T) = \frac{((1-\alpha)\lambda)^{y_0+y_1} B_1}{\exp((2-\alpha)\lambda)} \prod_{t=2}^{T} \frac{y_{t-1}!(1-\alpha)^{y_{t-1}+y_t}\lambda^{y_t} B_t}{\exp((1-\alpha)\lambda)} \quad (7.45)$$

The parameters α, λ, and y_0 can be estimated by maximizing the corresponding likelihood. The starting value problem, which is the more severe the shorter the time series, is discussed in detail in Ronning and Jung (1992). Brännäs (1995a) shows, how the INAR(1) model can be extended in order to include explanatory variables.

Example

Kennan (1985) analyses the frequency of, and duration between, contract strikes in the United Kingdom. The observations are from January 1968 to December 1976. The empirical mean of the series is 5.5, the empirical variance 13.4. The empirical overdispersion indicates that an INAR(1) process with Poisson marginals cannot be appropriate since this would require equidispersion. The actual time series is plotted in Fig. 7.1.

To illustrate the method, Fig. 7.2 displays a simulated INAR(1) process for $\alpha = 0.5$. The starting value corresponds to the observed value $y_0 = 5$ and the expected value of the marginal distribution is equal to the empirical mean. The simulation of the Poisson and binomial variables used algorithms from Kennedy and Gentle (1980). The mean and variance of the simulated time series are 5.6 and 4.9, respectively. The series thus corresponds much closer to the postulated equality of mean and variance. Also, it is clearly a stationary process with mean reversion.

INAR(1) Process With Unobserved Heterogeneity

The model (7.41) can be extended to allow for negative binomial marginals, and thus unoabserved heterogeneity and overdispersion McKenzie, 1986). Consider again the basic relationship

Fig. 7.1. Kennan's Strike Data

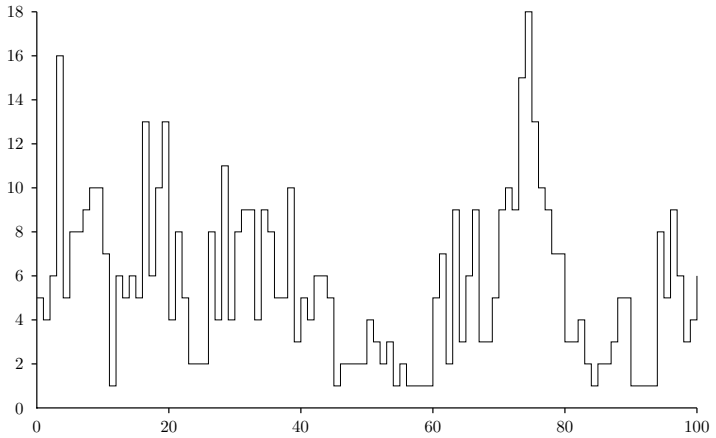
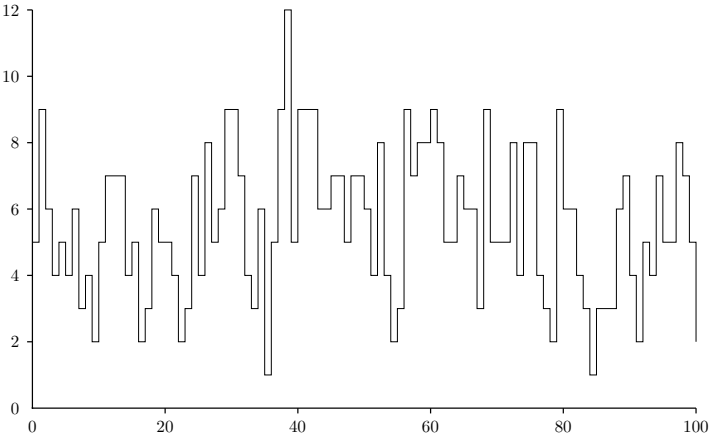


Fig. 7.2. Simulated INAR(1) Time Series for $\alpha = 0.5$



$$y_t = \alpha \circ y_{t-1} + \varepsilon_t$$

Assume that y_{t-1} has a negative binomial distribution with generic parameterization Negbin(δ, θ) (i.e., this is not a mean parameterization; rather, $\mathrm{E}(y_{t-1}) = \delta\theta$). In analogy to the Poisson case, one would be interested in a thinning operation $\circ$ that preserves the negative binomial distribution for $z_{t-1} = \alpha \circ y_{t-1}$. In the Poisson case we assumed that conditional on y_{t-1} and α, $z_{t-1} \sim$ binomial(y_{t-1}, α). McKenzie (1986) suggests that randomizing α through an indendent beta-distribution Be$(\gamma, \delta - \gamma)$ has a similar effect. In particular, $z_{t-1}|y_{t-1}, \gamma, \delta$ has a beta-binomial distribution and beta-binomial thinning results. It can be shown that the unconditional distribution of z_{t-1} is negative binomial Negbin(γ, θ). If, moreover, $\varepsilon_t \sim$ Negbin$(\delta - \gamma, \theta)$, an independent distribution, then it follows that the marginal distribution of y_t is Negbin(δ, θ).

Böckenholt (1999) discusses estimation of an INAR(1) process where unobserved heterogeneity is represented by a finite mixture, and where, conditional on the latent class, the process has all the standard properties of (7.41), including the Poisson marginals.

8

Bayesian Analysis of Count Data

The existing econometrics literature on count data models has largely ignored the Bayesian paradigm of inference. Likewise, in Zellner's (1971) influential book on Bayesian inference in econometrics, the Poisson regression model is not mentioned. The probable reasons for this neglect are computational complexities that in the past made the Bayesian analysis of count data models appear unattractive. However, increased computer power now allows for fast evaluation of posterior distributions by simulation methods. The basic approaches to Bayesian inference by simulation are discussed in this chapter.

In Bayesian econometrics the interest centers around the posterior distribution $\pi(\theta|y)$ which is a product of the likelihood function $f(y|\theta)/f(y)$ and a prior distribution $g(\theta)$

$$\pi(\theta|y) = \frac{f(y|\theta)g(\theta)}{f(y)} \tag{8.1}$$

where

$$f(y) = \int_{\Theta} f(y|\theta)g(\theta)\mathrm{d}\theta \tag{8.2}$$

does not depend on θ and is a normalizing constant of integration, also called *marginal likelihood*. This constant is often difficult to evaluate, and so is, as a consequence, the posterior distribution. The standard approach is to omit the normalizing constant and write

$$\pi(\theta|y) \propto f(y|\theta)g(\theta) \tag{8.3}$$

If the right hand side is the kernel of a known distribution, the normalizing constant can be infered from there. Alternatively, recent simulation based methods do not require an evaluation of the normalizing constant at all and thus are much more versatile.

In contrast to classical inference, Bayesian methods condition on the data and model the parameter as a random variable. While much of the debate on the relative merits of the Bayesian over the frequentist approaches has been

cast in philosophical terms, some of the recent literature has shifted the focus of the debate towards practical aspects: using recent simulation methods, the Bayesian approach can provide relatively simple solutions in models where frequentists methods fail, or at best, are difficult to implement. More on this below.

8.1 Bayesian Analysis of the Poisson Model

A standard result of a closed form posterior distribution exists for the Poisson model without covariates. Suppose $\{y_i\}, i = 1, \ldots, n$ is a random sample from a Poisson distribution with mean λ, and that the prior distribution of λ is a gamma distribution with parameters $\alpha \geq 0$ and $\beta \geq 0$. The gamma distribution is the conjugate prior for the Poisson parameter, and

$$\begin{aligned} g(\lambda|y) &\propto \left(\prod_{i=1}^{n} e^{-\lambda}\lambda^{y_i}\right) \frac{\alpha^{\beta}}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\lambda\beta} \\ &\propto e^{-\lambda(\beta-n)}\lambda^{\alpha+n\bar{y}-1} \end{aligned} \tag{8.4}$$

Hence, the posterior distribution of λ is a gamma distribution with parameters $\tilde{\alpha} = \alpha + n\bar{y}$ and $\tilde{\beta} = (\beta + n)$. Recall that the mean of the prior gamma distribution is given by $\mathrm{E}_0(\lambda) = \alpha/\beta$. Therefore, the posterior mean $\tilde{\alpha}/\tilde{\beta}$ can be written as

$$\begin{aligned} \mathrm{E}_{\pi}(\lambda|y, \alpha, \beta) &= \frac{\alpha + n\bar{y}}{\beta + n} \\ &= \frac{\beta}{\beta + n}\mathrm{E}_0(\lambda) + \frac{n}{\beta + n}\bar{y} \end{aligned}$$

The Poisson-gamma model is an example for a common result in Bayesian statistic, namely that the posterior mean is a weighted average of prior mean and sample mean. The weight given to the sample mean is an increasing function of the number of observations.

No conjugate prior exists for the $(k \times 1)$ parameter vector β in the Poisson regression model where the likelihood is proportional to

$$L(\beta|y, x) \propto \prod_{i=1}^{n} \exp[-\exp(x_i'\beta)]\,[\exp(x_i'\beta)]^{y_i} \tag{8.5}$$

Even with a noninformative prior, this expression is not the kernel of any known parametric distribution for β. There are two solutions. One is the use of approximation methods as, for instance, in Albert and Pepple (1989). The other is the evaluation of the exact posterior distribution using simulation methods. Consider approximation first. Let $\hat{\beta}$ be the mode of the posterior density, i.e., the maximum likelihood estimator. If the logarithm of this density is expanded in a second-order Taylor's series expansion around $\hat{\beta}$, we obtain

$$\ln \pi(\beta, y, x) \approx \ln L(\hat{\beta}) - \frac{1}{2}(\beta - \hat{\beta})' H (\beta - \hat{\beta}) \tag{8.6}$$

where H is minus the expected Hessian matrix evaluated at $\hat{\beta}$. Thus, the posterior of β is approximately multivariate normal with mean $\hat{\beta}$ and covariance matrix H^{-1}.

Next, assume a normal prior for β

$$g(\beta) = \phi(\beta | \beta_0, B_0^{-1}) \tag{8.7}$$

where β_0 denotes the prior mean, and B_0^{-1} the prior precision matrix, the inverse of the covariance matrix. That is,

$$g(\beta) \propto \exp[-1/2(\beta - \beta_0)' B_0^{-1} (\beta - \beta_0)]$$

In order to impose a reasonably vague prior, it is common to let $\beta_0 = 0$ and $B_0^{-1} = 10^{-2} I_k$, where I_k is the k-dimensional identity matrix. The posterior density $\pi(\beta|y)$ is then proportional to

$$\prod_{i=1}^{n} \exp[-\exp(x_i'\beta)] \, [\exp(x_i'\beta)]^{y_i} \exp[-1/2(\beta - \beta_0)' B_0^{-1} (\beta - \beta_0)] \tag{8.8}$$

and not in standard form.

Again, an approximation method could be used. Instead, in order to obtain exact results, we now consider simulation of the posterior density by the Metropolis-Hastings (MH) algorithm (See Chib and Greenberg, 1995). This is a special case of Markov chain Monte Carlo simulation. For a given target density $f(\psi)$, the MH algorithm is defined by

(1) a proposal density $q(\psi^\dagger|\psi)$ that is used to supply a proposal value $\psi^\dagger$ given the current value ψ, and
(2) a probability of move that is defined as

$$\alpha(\psi, \psi^\dagger) = \min \left\{ \frac{f(\psi^\dagger) q(\psi|\psi^\dagger)}{f(\psi) q(\psi^\dagger|\psi)}, 1 \right\}. \tag{8.9}$$

Hence, if $f(\psi^\dagger) q(\psi|\psi^\dagger) > f(\psi) q(\psi^\dagger|\psi)$ the chain moves to $\psi^\dagger$. Otherwise, it moves with probability $0 < \alpha(\psi, \psi^\dagger) < 1$. If rejected, the next sampled value is taken to be ψ.

For the MH algorithm to work efficiently, the choice of proposal density q is critical. Following Chib, Greenberg, and Winkelmann (1998), the proposal distribution for the Poisson regression model can be based on the mode $\hat{\beta}$ and curvature $V_\beta = [-H_\beta]^{-1}$ of $\ln \pi(\beta|y)$ where these quantities are found using a few Newton-Raphson steps with gradient vector

$$g_\beta = B_0^{-1}(\beta - \beta_0) + \sum_{i=1}^{n} [y_i - \exp(x_i'\beta)] x_i$$

and Hessian matrix

$$H_\beta = -B_0^{-1} - \sum_{i=1}^{n} \exp(x_i'\beta) x_i x_i'$$

The proposal can be obtained by reflecting the current value around the modal value $\hat{\beta}$ and then adding a Gaussian increment with variance τV_β (τ is a scalar that is adjusted in trial runs in order to obtain acceptance rates between 40 and 60 percent). The resulting proposal density is

$$q(\beta, \beta^\dagger | y) = \phi(\beta^\dagger | \hat{\beta} - (\beta - \hat{\beta}), \tau V_\beta)$$

To draw from the proposal density, we simply compute

$$\beta^\dagger = \hat{\beta} - (\beta - \hat{\beta}) + \tau \text{chol}(V_\beta)' \text{rndn}(k, 1)$$

where $P = \text{chol}(V_\beta)$ gives the Cholesky (upper-triangular) decomposition of V_β such that $V_\beta = P'P$, and $\text{rndn}(k, 1)$ is a $(k \times 1)$ vector of standard normal pseudo-random numbers.

Finally, the probability of move is given in terms of the ratio of density ordinates

$$\alpha(\beta, \beta^\dagger | y) = \min \left\{ \frac{\pi(\beta^\dagger | y)}{\pi(\beta | y)}, 1 \right\}, \tag{8.10}$$

since the proposal density is symmetric in $(\beta, \beta^\dagger)$ and hence cancels. In practice, the algorithms goes through a large number of iterations: 2,000 or 10,000 are some common values. In addition, it is recommended to preceed the proper sampling from the posterior by a burn-in phase of a given number of iterations (500, say), in order to reduce the influence of arbitrary starting values and let the algorithm move to its main area of support. Finally, the posterior sample can be analysed in order to report any distributional characteristics of choice, such as (posterior) mean, standard deviation, median, percentiles, or credibility intervals.

Discussion

The example of posterior simulation in the Poisson regression model illustrates well the potential of the method. Part of the simplicity arises since an evaluation of the normalizing constant is not required. This method has some direct additional benefits. For instance, inequality constraints on parameters can be imposed without problem: if sampled values fall within the inadmissable area, they are simply dropped. Likewise, it is also very simple to simulate the posterior distribution of a (possibly complicated) function of the parameters. In contrast to maximum likelihood, where the invariance property applies to the modal estimates, but standard errors need to be derived using asymptotic properties and the delta rule, the simulation approach immediately provides the full posterior distribution of the function, including correct standard errors, percentiles, etc. The approach is easily extended to other prior distributions. Finally, modified Poisson distributions (such as

truncated, censored) can be introduced by simply adjusting the likelihood function in (8.8).

8.2 A Poisson Model with Underreporting

In the Poisson regression model, the basic simulation tool was the Metropolis-Hastings step to draw from the posterior distribution of β. The power and versatility of Markov chain Monte Carlo can be substantially increased by combining MH-simulation with Gibbs sampling and data augmentation. For more detailed references on Markov chain Monte Carlo, see Chib and Greenberg (1996) and Gamerman (1997).

The following application to a Poisson model with underreporting illustrates these possibilites. Count data models with underreporting have been discussed in Chap. 6.5.2. The presentation here follows Winkelmann (1996b) who re-analysed the model in a Bayesian framework.

Let y_i^* denote the total number of events during a fixed time period T for individual i, and assume that the likelihood function $f(y_i^*|\beta)$ is of standard Poisson form, i.e.,

$$f(y_i^*|\beta) = \frac{\exp(-\exp(x_i'\beta))\exp(x_i'\beta)^{y_i^*}}{y_i^*!}$$

If y_i^* was observed, the algorithm of the previous section could be used to obtain the posterior distribution of β.

With underreporting, y_i^* is unobsered. We observe the number of reported events y_i which, conditional on y_i^*, is binomial distributed

$$f(y_i|y_i^*, p_i) = \frac{y_i^*!}{(y_i^* - y_i)!y_i!} p_i^{y_i}(1-p_i)^{y_i^*-y_i} \tag{8.11}$$

where p_i gives the individual probability of reporting an event. The structure of the model becomes more apparent once we write down the joint posterior distribution of β, p and y^*, where y^* has been included among the parameters, a case of data augmentation:

$$\pi(y^*, p, \beta|y, x) \propto f(y|y^*, p, \beta)f(y^*|\beta)g(\beta)g(p) \tag{8.12}$$

The following prior distributions g can be used:

$$g(\beta) \sim \phi(\beta_0, B_0^{-1})\,. \tag{8.13}$$

and

$$g(p) \sim \mathcal{U}(0,1) \tag{8.14}$$

where $\mathcal{U}(0,1)$ is the standard uniform distribution. The resulting joint posterior distribution of y_i^*, p_i, and β is then proportional to

$$\pi(y^*, p, \beta | y, x) \propto \prod_{i=1}^{n} \exp\{y_i^* x_i' \beta - \exp(x_i' \beta)\} \frac{p_i^{y_i} (1-p_i)^{y_i^* - y_i}}{(y_i^* - y_i)! y_i!} \times \exp(-1/2(\beta - \beta_0)' B_0 (\beta - \beta_0)) \tag{8.15}$$

While it is intractable to derive analytically the marginal posterior distributions for the parameters of interest from (8.15), the MCMC approach allows to simulate the joint posterior density. One could contemplate a direct "brute-force" simulation of (8.15) using the MH method described in the previous section. However, it will be problematic to obtain a suitable proposal density and the approach is likely to be costly and inefficient.

A superior algorithm is to sample the joint posterior by successively sampling through its full conditional distributions. This is also frequently refereed to as Gibbs sampling. The gains from Gibbs sampling are most evident when some of the full conditional distributions can be simulated from standard distributions, as is the case in this example. The three required full conditional distributions in this case are

$$[y^* | p, \beta, y, x], \ [p | y^*, \beta, y, x], \ [\beta | y^*, p, y, x]$$

Inspection of the joint posterior (8.15) reveals that the full conditional posterior of y^* is given by

$$\pi(y^* | \beta, p, y, x) \propto \prod_{i=1}^{n} \frac{[\exp(x_i' \beta)(1-p_i)]^{y_i^*}}{(y_i^* - y_i)!}$$

This is the kernel of a Poisson distribution shifted by y_i, with parameter $\lambda_i^* = \exp(x_i' \beta)(1-p_i)$. Pseudo random numbers from this distribution can be readily generated using the Knuth (1969) algorithm.

The full conditional distribution of p is of beta form, while the full conditional distribution of β (that depends on y^* only) is the same as the posterior distribution of the Poisson regression model and can thus be simulated using the MH algorithm discussed in the previous section. This last step exactly reflects the effect of data augmentation: data augmentation replaces unobserved values by simulated values and thus restores the model in standard form.

The sampling process is initiated with values in the support of the posterior density. The algorithm runs in cycles through the three full conditional densities, where the conditioning values for the parameters are updated as soon as a more recent value becomes available. As for the MH algorithm, the simulated values after an initial burn-in phase are retained as a sample from the target joint posterior distribution. It also should be noted that the MH-step required in the simulation of β involves only one draw in each cycle of the Gibbs sampler.

8.3 Estimation of the Multivariate Poisson-Log-Normal Model by MCMC

Recall the multivariate Poisson-log-normal model presented in Chap. 7.1.4. Conditionally on a $(J \times 1)$ vector of individual and outcome specific random effects $\varepsilon_i = (\varepsilon_{i1} \cdots \varepsilon_{iJ})'$ the distribution of y_i is independent Poisson

$$f(y_i|\varepsilon_i) = \prod_{j=1}^{J} f_p(y_{ij}|\lambda_{ij} \exp(\varepsilon_{ij})) \tag{8.16}$$

Moreover

$$f(\varepsilon_i|\Omega) = \phi_J(-0.5\text{diag}\Omega, \Omega) \tag{8.17}$$

where ϕ_J is the J-variate normal density with covariance matrix Ω. The details of this model structure were discussed before. For the purposes of the present chapter, the important aspect is that the likelihood function requires the evaluation of a J-variate integral of the Poisson distribution with respect to the distribution of ε_i

$$f(y_i|\lambda_i, \Omega) = \int \prod_{j=1}^{J} f_p(y_{ij}|\lambda_{ij}, \varepsilon_{ij})\phi(\varepsilon_i| - 0.5\text{diag}\Omega, \Omega)\text{d}\varepsilon_i$$

and hence estimation by maximum likelihood may not be possible for large J. Chib and Winkelmann (2001) show how the joint posterior distribution of $\psi = (\beta, \Omega)$ can be obtained using MCMC for arbitrary J (An application of the Gibbs sampler to a univariate Poisson-log-normal model, i.e., for $J = 1$, is provided by Plassmann and Tideman, 2001). Suppose that the parameters (β, Ω) independently follow the prior distributions

$$\beta \sim N(\beta_0, B_0^{-1}), \quad \Omega^{-1} \sim \text{Wish}(\nu_0, R_0),$$

where (β_0, B_0, v_0, R_0) are known hyperparameters and $\text{Wish}(\cdot, \cdot)$ is the Wishart distribution with ν_0 degrees of freedom and scale matrix R_0.

Using data augmentation for the random effects ε, the parameters can be blocked as ε, β, and Ω after which the joint posterior is simulated by recursively sampling the full conditional distributions

$$[\varepsilon|y, \beta, \Omega]; \quad [\beta|y, \varepsilon]; \quad [\Omega^{-1}|\varepsilon]\,, \tag{8.18}$$

using the most recent values of the conditioning variables at each step.

In order to sample ε from the target density

$$\pi(\varepsilon|y, \beta, \Omega) = \prod_{i=1}^{n} \pi(\varepsilon_i|y_i, \beta, \Omega)$$

consider the i-th target density

$$\pi(\varepsilon_i|y_i, \beta, \Omega\,) = c_i\phi(\varepsilon_i|\Omega) \prod_{j=1}^{J} \exp(-\tilde{\lambda}_{ij})\tilde{\lambda}_{ij}^{y_{ij}} \tag{8.19}$$

where $\tilde{\lambda}_{ij} = \exp(x'_{ij}\beta_j + \varepsilon_{ij})$. Draws from this conditional density can be obtained using the MH-algorithm. The proposal density is taken to be multivariate-t with parameters that are tailored to those of the target $\pi(\varepsilon_i|y_i, \beta, \Omega)$. Let $\hat{\varepsilon}_i = \arg\max \ln \pi(\varepsilon_i|y_i, \beta, \Omega)$ and $V_{\varepsilon_i} = (-H_{\varepsilon_i})^{-1}$ be the inverse of the Hessian of $\ln \pi(\varepsilon_i|y_i, \beta, \Omega)$ at the mode $\hat{\varepsilon}_i$. Then, the proposal density is given by $q(\varepsilon_i|y_i, \beta, D) = f_T(\varepsilon_i|\hat{\varepsilon}_i, V_{\varepsilon_i}, \nu)$, a multivariate-$t$ density with ν degrees of freedom (where ν is a tuning parameter).

The sampling of β, conditional on ε, follows the standard Poisson procedure discussed in Chap. 8.1. Finally, Ω^{-1} is sampled by MH from a target density proportional to

$$\pi = f_W(\Omega^{-1}|v_0, R_0^{-1}) \prod_{i=1}^{n} \phi(\varepsilon_i| - 0.5\text{diag}\Omega, \Omega) \tag{8.20}$$

where f_W denotes again the Wishart density. See Chib and Winkelmann (2001) for further details and applications.

8.4 Estimation of a Random Coefficients Model by MCMC

With panel data, parameter heterogeneity becomes a possible issue. While most of the literature has dealt with intercept heterogeneity, this is not an intrinsic constraint. In fact, among linear model, the class of random coefficient models explicitly introduce heterogeneity for all regression coefficients of the model, including the slopes (see, e.g., Greene, 1993). Unfortunately, few of the known results for the linear model carry over to non-linear models such as the Poisson model. In fact, such a class of count data models may be intractable using maximum likelihood based methods. However, Chib, Greenberg, and Winkelmann (1998) show how a random coefficients model can be analysed in a Bayesian framework using MCMC. As in the previous chapter, data augmentation and Gibbs sampling are the key elements of the algorithm.

The main elements of the model are as follows: The distribution of $y_i = (y_{i1} \cdots y_{iT})'$ conditional on (b_i, β) is assumed to be independent Poisson

$$f(y_i|\beta, b_i) = \prod_{t=1}^{T} \frac{\exp(-\tilde{\lambda}_{it})\ \tilde{\lambda}_{it}^{y_{it}}}{y_{it}!} \ .$$

where

$$\tilde{\lambda}_{it} = \exp(x'_{it}\beta + w'_{it}b_i)$$

β are fixed coefficients and b_i are random coefficients. It should be stressed that w is not a subset of x. In a pure random coefficients model, all explanatory variables are part of w and x does not appear in the conditional mean function.

The random coefficients have a multivariate normal distribution

$b_i \sim N_q(\eta, \Omega)$

The (unknown) vector η represents the corresponding fixed effects. The joint density of $y = (y_1, ..., y_n)$ conditional on β, η and Ω (but unconditional on b_i) is given by

$$f(y|\beta, \eta, \Omega) = \prod_{i=1}^{n} f(y_i|\beta, \eta, \Omega) \tag{8.21}$$
$$= \prod_{i=1}^{n} \int \prod_{t=1}^{T} f(y_{it}|\beta, b_i)\, \phi(b_i|\eta, \Omega)\, d\, b_i$$

If one were to proceed with maximum likelihood estimation (8.21) could be maximized with respect to β, η and D as long as there are few random coefficients and hence a low-dimensional integral structure.

Here, the interest is in the joint posterior. It is given by

$f(\beta, \eta, b, \Omega|y) \propto f(y|\beta, b) f(\beta) f(b|\eta, \Omega) f(\eta) f(\Omega)$

with priors

$\beta \sim N_k(\beta_0, B_0^{-1})$, $\eta \sim N_k(\eta_0, M_0^{-1})$, $\Omega^{-1} \sim \text{Wish}(v_0, R_0)$,

where Wish is the Wishart distribution. The joint posterior can then be rewritten as (up to a proportionality constant)

$$\phi(\beta|\beta_0, B_0^{-1}) \phi(\eta|\eta_0, M_0^{-1})\, f_W(\Omega^{-1}|v_0, R_0) \prod_{i=1}^{n} f(y_i|\beta, b_i)\, \phi(b_i|\eta, \Omega)$$

To simulate the posterior distribution by Markov Chain Monte Carlo, the following simple conditional structure is used.

- $[\beta|y, b, \Omega]$ is proportional to $[\beta] \times [y|\beta, b]$ because the conditional density of the observable data is independent of Ω. Its density is proportional to

$$\exp\left(-\frac{1}{2}(\beta - \beta_0)' B_0 (\beta - \beta_0)\right) \prod_{i=1}^{n} \prod_{t=1}^{T} \exp(-\tilde{\lambda}_{it}) \tilde{\lambda}_{it}^{y_{it}}$$

- The conditional distribution $[\eta|y, \beta, b, \Omega]$ is proportional to $[\eta] \times [b|\eta, \Omega]$, independent of y and β, and is given by

  $N_q(\hat{\eta}, M_1^{-1})$

  where $\hat{\eta} = M_1^{-1}(M_0\eta_0 + \sum_{i=1}^{n} \Omega^{-1} b_i)$ and $M_1 = (M_0 + n\Omega^{-1})$.
- The conditional distribution $[b|y, \beta, \Omega]$ factors into a product of conditionally independent distributions

$$[b|y, \beta, \eta, \Omega] = \prod_{i=1}^{n} [b_i|y_i, \beta, \eta, \Omega]$$

  where $[b_i|y_i, \beta, \eta, \Omega]$ is proportional to $[b_i|\eta, \Omega]$ $[y_i|\beta, b_i]$ with density given by

$$\exp\left(-\frac{1}{2}(b_i-\eta)'\Omega^{-1}(b_i-\eta)\right)\prod_{t=1}^{T}\exp(-\tilde{\lambda}_{it})\tilde{\lambda}_{it}^{y_{it}}$$

- The conditional distribution $[\Omega^{-1}|y,\beta,\eta,b]$ is proportional to $[\Omega^{-1}]\,[b|\eta,\Omega]$ wich can be shown to be of Wishart form

$$\text{Wish}\left(n+v_0,\left[R_0^{-1}+\sum_{i=1}^{n}(b_i-\eta)(b_i-\eta)'\right]^{-1}\right)$$

Except for the conditional distributions of η and Ω^{-1} these are unknown distribution families that are simulated using the Metropolis-Hastings algorithm. Successively sampling the distributions

$$[\beta|y,\eta,b,\Omega]\rightarrow[\eta|y,\beta,b,\Omega]\rightarrow[b|y,\beta,\Omega]\rightarrow[\Omega^{-1}|y,\beta,b]$$

generates then a Markov Chain sample that converges to the target density.

9

Applications

9.1 Accidents

The study of accident rates is one of the earliest applications of count data analysis: Bortkiewicz (1898) established that the annual number of deaths from mule kicks in the Prussian army during the period 1875-1894 could be well described by a Poisson distribution with mean of 0.7 corps per year. Although later studies have shifted their interest to different causes of death, the same principles apply. Accidents, by their very definition, come possibly closest to the notion of randomness required for processes such as the Poisson process.

For instance, Evans and Graham (1990) investigate the effect of child restraint use legislation on child mortality in car accidents, using state level data. The risk set is defined as total vehicle miles traveled (per year and state, in thousands). Michener and Tighe (1992) in a similar analysis investigate the effects of speed limits, minimum drinking age, and mandatory seatbelt laws on the number of fatal accidents. They estimate a number of models, where they control (as offset) either for vehicle miles traveled, for the number of registered drivers or for the number of registered vehicles as indicators of the size of the risk set. More recently, Hahn and Prieger (2006) and Prieger and Hahn (2007) estimate the effect of cell phone use while driving on the number of car accidents per quarter, using a survey of individual-level data on cell phone usage and driving patterns. Rose (1990) analyzes the determinants of air-traffic incidents per number of scheduled departures per year (in thousands), and Kahn (2005) studies the number of deaths from natural disaster.

As discussed in Chap. 3.1.5 these scale variables can be either included in the regression as logarithmic offset with unit coefficient (as in Rose, 1990) or the coefficient can be given free for estimation (as in Evans and Graham, 1990, and Michener and Tighe, 1992). In the former case, the interpretation of the regression function $\exp(x_i'\beta)$ is in terms of an accident rate.

The focus of the above studies, and numerous others of similar kind, is on evaluating public policy: how successful was past (traffic) safety legislation

in reducing the number of accidents. Frequently, such studies use aggregate data. Another type of analysis has pursued the insurance aspect of accidents. For instance, based on individual level data, it would be interesting to estimate the actuarial risk associated with insuring a certain individual with given characteristics and record. Although the data are different, the employed methods are essentially the same. For instance, Dionne and Vanasse (1992) use driving records and Negbin II regressions in order to predict individual claim frequencies, taking into account individuals characteristics and driving histories. Dionne et al. (1994, 1995) extend this analysis by controlling for various aspects of the driver's medical condition.

Applications of count data models to accident numbers are by no means limited to those involving transport. Feinstein (1989) uses pseudo-maximum likelihood Poisson regressions in order to model the number of "incidents" at U.S. nuclear power plants. Bauer et al. (1998) are interested in the determinants of the number of workplace accidents using firm level data for Germany. Ruser (1991, 1993) studies the number of workdays lost due to occupational injury.

9.2 Crime

Grogger (1990a) analyzes the short-term deterring effect of capital punishment. To be specific, he investigates whether the daily number of homicides in California in the early 1960s were lower on days that followed an execution (or on days leading up to one, as they were usually reported in the newspapers with a lead), than on other days. Using Poisson and negative binomial regression models, the null-hypothesis of no effect cannot be rejected.

Kelly (2000) analyzes the effects of inequality on the number of property and violent crimes. Data are at the U.S. county level for the year 1990. The employed technique is Poisson quasi-likelihood estimation. Kelly finds that for violent crime, there is a large impact of inequality even after controlling for the effects of poverty, race and family composition. There is a negative effect of policy activity on crime that persists once the possible endogeneity of police activity is accounted for in an instrumental variable GMM procedure.

Finally, Plassmann and Tideman (2001) estimate generalized Poisson models to examine the dynamic effects of right to carry laws on reported homicides, rapes, and robberies, again using data for the U.S. with variation over time and across U.S. states. They conclude that such laws appear to have statistically significant deterrent effects on the numbers of reported murders, rapes, and robberies.

9.3 Trip Frequency

A natural application of count data modeling arises when one is interested in finding out what determines the number of trips taken by a person (or

household) over a specific period of time. Such data occur, for example, in empirical studies in the fields of environmental economics and regional economics. In the former, trip frequency can be used to estimate the value that a recreational site, such as a hiking trail, a swimming beach, or a river for fishing provides to its users. Such an estimate can enter into a cost-benefit analysis when elimination of such a recreational site is considered. In regional economics, and urban planning in particular, one is interested in the number of trips to a particular shopping site, and how this number is affected by the distance, the characteristics of the shopping site, and the location and attributes of alternative sites in the region.

The demand for recreational trips was analyzed, among others, by Creel and Loomis (1990), Ozuna and Gomez (1994, 1995), Shonkwiler and Shaw (1996), Gurmu and Trivedi (1996), Haab and McConnell (1996), and Hellström (2002, Chap. 4). The prevailing approach is the so-called "travel cost method". The goal of these models is to estimate a conventional downward sloping demand function. The 'quantity demanded' is the number of trips taken to a site during a given period of time, and the 'price' is the travel cost of reaching the site. Price variation derives from the fact that individuals live at different distances from the site. Those living nearby have lower cost and would be expected to undertake more trips. Formally, let y_i, the number of trips to a single site by individual (or household) i, have a count data distribution with

$$\mathrm{E}(y_i) = \exp(\beta_1 tc_i + \beta_2 d_i)$$

where tc_i are the cost per trip and d_i are various socio-economic characteristics, including income. The parameters β_1 and β_2 can be estimated in the usual way, once the model has been specified and a random sample of potential users has been surveyed. If the sample is taken at the site, corresponding adjustments have to be made to account for the endogenous sampling (see Chap. 5.1.2).

Presumably, β_1 is negative. The consumer surplus for individual i can then be calculated as the integral under the demand function from tc_i to tc_i^*, where tc_i^* is the cost that would choke of demand, i.e., reduce the demand to (approximately) zero. In the log-linear form used here, where demand approaches zero only in the limit, we have

$$\begin{aligned} cs_i &= \int_{tc_i}^{\infty} \exp(\beta_1 tc_i + \beta_2 d_i) d\, tc_i \\ &= \frac{1}{\beta_1} \exp(\beta_1 tc_i + \beta_2 d_i)\Big|_{tc_i}^{\infty} \\ &= -\frac{\mathrm{E}(y_i)}{\beta_1} \end{aligned}$$

This would be the access value for individual i that would be lost if the site was closed. Aggregation over all affected individuals would give the overall consumer surplus lost that should be counted among the cost of a site-closure.

Extensions of this model recognize that the "use value" of one site may be affected by the availability and location of other similar sites, and their respective attributes. This is addressed in so-called multiple site models. Also, one has to recognize that the value of a site may not be limited to its use-value. Even non-users might attach a values to the availability of a site, and different methods are required to determine the "non-use value".

The second area where trip frequencies have been subject to count data modeling in the past is for the number of trips to a number of alternative retail locations (Flowerdew and Aitkin, 1982, Okoruwa, Terza, and Nourse, 1988, Barmby and Doornik, 1989). Again, distance from the site is an important explanatory variable. In addition, the effect of the size of the retail location (or shopping mall) is of interest, following the notion of a gravity model that the attractive force of a location is directly proportional to its size and inversely proportional to the distance or travel time to that location. Estimation of such a model using Poisson or negative binomial regression models is straightforward. In these models, one can also control for further socio-economic individual specific variables that may affect shopping behavior. A further application of such gravity models using count data is the modeling of spatial flows (for instance of people) as in Smith (1987).

9.4 Health Economics

Count data models have a wide applicability in health economics and, more generally, in health sciences. A number of separate application areas can be distinguished.

Firstly, count data models are used to model the frequency or intensity of a health problem. For instance, using individual level patient records from a controlled experiment, Diggle, Liang and Zeger (1995) and Chib, Greenberg and Winkelmann (1998) estimate the effect of a drug treatment on the number of epileptic seizures over a given period. Mullahy and Portney (1990) use survey data from the 1979 National Health Interview Survey to estimate the effect of smoking on the number of days of respiratory illness during a two-week recall period. Jorgensen et al. (1999) study the number of emergency room visits for respiratory diseases, again using count data regression models. Böhning et al. (1999) provide an epidemiological application to the number of teeth affected by caries.

A second, related area of research concerns the occurrence and incidence of patterns of behavior that are perceived or known to be potentially harmful and "unhealthy". One such issue, related to public health, is the freqency of sexual intercourse of teenagers (for instance, Moffatt and Peters, 2000). Other examples are the determinants of cigarette or alcohol consumption (for instance, Yen, 1999, and Kenkel and Terza, 2001).

A third area of application comprises the utilization of health services. Often, household survey data provide information on variables such as the

number of doctor consultations over the previous three or twelve months. Such data are relatively inexpensive to obtain and, although lacking information on the cost of the service or the associated diagnosis, they constitute a good first indicator of health care utilization. In fact, it has been argued that the number of doctor consultations can be used as a proxy for health status per se (Møller Danø, 1998). Early papers include Cameron et al. (1988) and Pohlmeier and Ulrich (1995). This type of analysis has seen a veritable explosion of activity in recent years. See, for example, Vera-Hernandez (1999), Schellhorn (2001), Jimenéz-Martin, Labeaga and Martinez-Granado (2002), Doorslaer, Koolman and Jones (2004), and Winkelmann (2004a,b, 2006).

The number of doctor consultations is not the only indicator of this sort. Depending on data availability, a distinction can be made between visits to a general practitioner, to a specialist, to a dentist (Melkersson and Olsson, 1999, or to a non-doctor health professional (Gurmu and Elder, 1998). Alternatively, other aspects of health utilization can been studied, such as the number of hospital outpatient department visits (Freund et al., 1996), the number of emergency room visits (Freund et al., 1996), the number of hospital inpatient days (Freund et al., 1996, Geil et al., 1997), or the amount of home care received by the elderly (Gameren and Woittiez, 2002).

In related types of research using count data methods, Grootendorst (1995) studies the usage of prescription drugs and Jensen, 1987, studies the discovery of new drugs. Finally, one can count among the health related studies those that address worker absenteeism due to illness, or worker absenteeism in general (Vistnes, 1997, Winkelmann, 1999, Barmby, Nolan and Winkelmann, 1999).

Studies of health outcomes often attempt to estimate the effect of a treatment, controling for the general effects of socio-economic characteristics (e.g., age, sex, ethnicity, labor force status). The paper by Melkersson and Olsson (1999) is an example of a treatment study as it estimates the effect of preventive dental care during childhood and adolescence on dental health (measured by the number of visits to a dentist) as an adult. Møller Danø (1999) estimates the effects of unemployment on health, contributing to the wider literature on the "social cost of unemployment".

Studies of health utilization frequently draw their motivation from microeconomic theory and the analysis of demand and supply in the market for health. Three leading areas of interest are the sign and size of the income effect, the role of health insurance (as health insurance implicitly determines the relative price of health services and thereby the substitution effect), and the effect of supply (physician density) on demand, if any. Examples in this literature include Grootendorst (1995), Gerdtham (1997), and Winkelmann (2004a,b). Grootendorst finds, using Canadian data, that the removal of co-payments for prescription medicines increased the utilization of prescription drugs. Winkelmann (2004a,b) reports that a 1997 German health care reform that increased the co-payments by up to 200 percent, reduced the demand for doctor visits by around 10 percent.

Geil et al. (1997) establish that private insurance has no effect on the number of admissions to a hospital in Germany, while Cameron et al. (1988) find for Australia that the usage of health services increased with coverage of the insurance policies. Freund et al. (1996) use differences-in-differences to estimate the impact of Medicaid changes where fee-for-service coverage is replaced by managed care. They report significant reductions in the hospital outpatient usage under managed care. Finally, Pohlmeier and Ulrich (1995) find some evidence for supplier-induced demand: a higher density of physicians leads to increased usage, as measured by the number of doctor consultations.

Count data based empirical research in the health area typically encounters the full array of methodological problems discussed in this book. Most research is based on single outcome measures but sometimes several outcomes are modelled jointly. For instance, Riphahn, Wambach and Million (2003) use the multivariate Poisson-log-normal model to jointly estimate the determinants of the number of visits to a doctor and the number of visits to a hospital. Their model allows in addition for individual specific random effects, as their application is based on panel data. See also Gurmu and Elder (1998).

Of course, as in other applications, overdispersion is common in health data. In many cases, the excess of zeros is so great that it cannot be accommodated by the negative binomial distribution or similar single-index models. As a consequence, two-part models, and hurdle models in particular, have become the method of choice in empirical applications (Pohlmeier and Ulrich, 1995). One interesting aspect of those models is that they sometimes may be given a structural interpretation. For instance, as far as the number of doctor visits during a given period of time is concerned, the hurdle part can explain the decision to contact a general practitioner (GP), i.e., the onset of a sickness spell. Once a GP has been contacted, further referrals follow a different process that is to a considerable degree determined by the decisions of the GP rather than the individual. Recently, Santos Silva and Windmeijer (1999) have pointed out that hurdle models are unnecessarily restrictive as they allow at most one sickness spell during the given period of time. Instead, one can use the more general framework of compound count data distributions to jointly model the number of sickness spells and the number of referrals per spell.

In an application of zero inflated models, Gameren and Woittiez (2002) estimate the determinants of the demand for home care by the elderly. The dependent variable is the number of shifts of home care received per week. The zero inflated model has an interesting interpretation in this context, because in the case considered by the authors, the Netherlands in 1996, demand was rationed as was evidenced by substantial waiting lists. Hence, there are two types of non-users, those without demand and those with demand but rationed by waiting lists. The model in principle allows to disentangle these two effects.

Another problem in empirical health economics is the potential endogeneity of explanatory variables. This problem has moved to the forefront of recent research. Solutions have been proposed among others by Freund et al. (1996), Mullahy and Portney (1990), and Windmeijer and Santos Silva (1997) with

particular applications in health economics in mind. With endogeneity, inconsistency will arise and the desired interpretation of the estimates as a causal relation becomes inadmissible. An example for the problem of endogeneity is given by the effect of insurance coverage. If individuals can choose their coverage then economic theory predicts a process of "adverse selection". With imperfect and asymmetric information individuals whose high health risk is known to themselves but not to others, including the insurer, will choose the high coverage insurance policy. A related problem is that of "moral hazard": high coverage may lead to negligent behavior and reduced preventive care on the part of the individual.

In either case, the observed insurance effect will not necessarily measure the causal behavioral response of insurance on health utilization. In order to address this problem, instrumental variable and switching regression estimators have been proposed. Freund et al. (1996) use state variation in changes to Medicaid laws as an instrument. Schellhorn (2002), using data for Switzerland, estimates the effect of choosing a higher deductible on the number of doctor visits. Clearly, those with low expected use will benefit from selecting a high deductible. The results indicate that the effect of choosing a higher deductible is overestimated when this self selection is not controlled for. Depending on the method, all of the observed difference in utilization can be explained by self-selection.

9.5 Demography

The main application of count data models in demography is the analysis of individual fertility, as measured by the number of children ever born or the number of children living in a household. Examples for recent applications are Nguyen-Dinh (1997) and Al-Qudsi (1998a, 1998b). The *Journal of Population Economics* devoted a symposium to fertility studies using count data models (Winkelmann and Zimmermann, 2000).

Modeling fertility produces a number of interesting methodological issues. These include, in no particular order, the frequent presence of underdispersion (Winkelmann and Zimmermann, 1994, Winkelmann, 1995), the influence of infertility and social norms (as opposed to individual choice based on economic factors), and the question of how to account for the fact that women may not have yet completed their childbearing age (Caudill and Mixon, 1995, McIntosh, 1999).

A number of approaches have been taken in order to deal with incomplete fertility. The most radical one is to consider older women only, for instance those aged 45 or older, in order to bypass the problem. Examples are Winkelmann (1995) and Mayer and Riphahn (2000). This "method" has a couple of drawbacks, however. First, the omission of data on the current child-bearing generation generates a substantial lag in the collection of evidence on fertility patterns. This becomes more of a problem if fertility behavior is rapidly

changing over cohorts. Second, the method cannot be used if the number of children is based on household composition data (such as in Famoye and Wang, 1997, and Kalwij, 2000), as children typically leave the household once they reach adulthood.

Instead, one can include a variable such as age, or age-at-marriage, or the number of fertile years, as logarithmic offset in the regression (see Chap. 3.1.5). Alternatively, one can consider models where the number of children observed for women with incomplete fertility is interpreted as a lower bound of completed fertility. A corresponding censored probability model is relatively simple to establish. Such models due to Caudill and Mixon (1995) and McIntosh (1999) were discussed in Chap. 5.1.1. The two contributions differ in the way "completion status" is determined. In Caudill and Mixon it is based on age whereas in McIntosh it is based on an additional survey question on desired fertility.

When modeling the determinants of fertility there are strong reasons to believe that the standard assumption of a homogeneous exponential mean function is violated. These include the possibility of infertility (i.e., the outcome of zero children that results from processes other than choice), and the potential influence of social norms. For instance, in many societies, to have an only child is considered to be socially undesirable whereas to have two children is considered desirable. That "zeros" are different can also be seen in aggregate data. For instance, Santos Silva and Covas (2000) point out that in developed countries the average number of children per couple has fallen while the percentage of childless couples has remained relatively stable.

The offshoot of these considerations is that the homogeneity assumption underlying the exponential mean function of the count data model may be wrong in which case the standard Poisson-based estimator is inconsistent. Thus, the literature has considered alternative data generating processes. Two recent contributions to this area of research include Santos Silva and Covas (2000) and Melkersson and Roth (2000). Both papers model completed fertility ((de-facto) married women aged 40 or older in Portugal and (de-facto) married women aged 45 or older in Sweden, respectively). The two papers make, however, different assumptions on the processes that give extra weight to the outcomes zero, one or two.

Santos Silva and Covas combine a hurdle-at-zero model with inflation (or deflation) at one for the positive count data part.

$$f(y_i|x_i) = \begin{cases} g_1(0|x_i) & \text{for } y_i = 0 \\ (1 - g_1(0|x_i))(\omega + (1-\omega)g_2(1|x_i)) & \text{for } y_i = 1 \\ (1 - g_1(0|x_i))(1-\omega)g_2(y_i|x_i) & \text{for } y_i = 2, 3, 4, \ldots \end{cases}$$

Specifically, they assume that $g_1(y|\beta_1)$ is a generalized Poisson distribution (see Chap. 2.6.2), $g_2(y|\beta_2)$ is a truncated-at-zero generalized Poisson distribution, and

$$\omega = \frac{(\theta - 1)g_2(y|\beta_2)}{1 - (1-\theta)g_2(y|\beta_2)}$$

In this way, the generalized Poisson distribution without hurdle and in-/deflation is obtained for $\beta_1 = \beta_2$ and $\theta = 1$.

Melkersson and Roth (2000) devise a model that inflates both the "zero" and the "two" outcomes. The zero-and-two inflated model has the following probability distribution function

$$f(y_i|x_i) = \begin{cases} \omega_0 + (1-\omega_0-\omega_2)g(0|x_i) & \text{for } y_i = 0 \\ \omega_2 + (1-\omega_0-\omega_2)g(2|x_i) & \text{for } y_i = 2 \\ (1-\omega_0-\omega_2)g(y_i|x_i) & \text{for } y_i = 1,3,4,\ldots \end{cases}$$

where $g(y_i|x_i)$ is a proper count data distribution. In principle, the ω's can be negative, representing a shortfall of zero's or two's relative to the base model, as long as some inequality restrictions are observed (e.g., $\omega_0 > g(0|x_i)(\omega_2 - 1)/(1 - g(0|x_i))$). Of course, if the ω's are modeled as a logit-function of covariates as in Melkersson and Roth (2000), then zero- or two-deflation is precluded.

The findings of Melkersson and Roth and Santos Silva and Covas cast doubt on the assumption of a homogeneous count process in these situations, and suggest a re-interpretation of the phenomenon of underdispersion that is so characteristic of completed fertility data. Here, underdispersion stems from differences between the various components of the model, rather than from a mere departure from the Poisson variance function. Hence, the earlier practice of modeling fertility using count data models with generalized variance function (for instance, Winkelmann and Zimmermann, 1994, and Wang and Famoye, 1997) might be misguided, as a violation of the mean function leads to inconsistent parameter estimates.

A research area of substantive interest is the dynamic interaction between child-bearing and employment status over the life-cycle. While count data are certainly less than ideal to address such simultaneity, the contribution by Kalwij (2000) offers substantial progress in that direction. He makes the identifying assumption that a woman's employment status remains unchanged after birth of the first child. This assumption is supported by some simple descriptive evidence for Dutch women.

Under this assumption, the simultaneous choice of having at least one child and employment can be modeled using cross-section data only in a bivariate probit or multinomial logit framework, whereas the number of children for those who have at least one child is modeled conditional on employment. An important finding, using data from a Dutch household survey, is that the effects of educational attainment on the observed fertility pattern runs via the effects of educational attainment on female employment status, which in turn significantly affects the fertility behavior of households. The direct effect of educational attainment on the presence and number of children is found to be relatively small.

Mayer and Riphahn (2000) and Atella and Rosati (2000) use standard count data models to address novel questions related to the determinants of fertility. Meyer and Riphahn analyze the fertility adjustment of Guestworkers

in Germany. In particular, they are interested in the effect of the variable "fertile years in Germany", which is, by assumption, the number of years between the age of 15 and the age of 40 that an immigrant woman has spent in Germany. Using individual level data on completed fertility from the German Socio-Economic Panel, the evidence favors an "assimilation" hypothesis (a gradual decline to the lower fertility levels of German-born women) over a "disruption" hypothesis (an initial drop in fertility below native levels with subsequent catch-up). It is interesting to note that contrary to the well known identification problem that arises in the study of earnings assimilation, fertility assimilation as defined by the authors in fact does allow to disentangle cohort and assimilation effects even from pure cross-section data. This is so because a given arrival cohort can differ at any point in calendar time in the number of fertile years spent in Germany (by virtue of differences in age at arrival in Germany).

Finally, Atella and Rosati (2000) build a model of fertility decisions in the context of a developing country where children are a means of intergenerational asset transfer. In such a model fertility does not only depend on the expected survival rate of children but also on the uncertainty associated with this survival rate. The empirical analysis using data from India shows that increased uncertainty leads to lower fertility levels.

9.6 Marketing and Management

Count data regressions become increasingly common in marketing and management as well. A prime example is the analysis of consumer behavior in studies that attempt to explain and predict purchase frequencies or amounts (for instance, Wedel et al. 1993, Robin, 1993, Ramaswamy, Anderson and DeSarbo, 1994, Brockett, Golden and Panjer, 1996). A related problem is that of modeling consumer brand choice (for instance, Gupta, 1988, Dillon and Gupta, 1996). In such models, the number of purchases of a certain brand is modeled conditional on the total number of purchases of a given item (over a year, say). The resulting model, similar to the conditional likelihood approach of the fixed effects Poisson model, is of a multinomial logit form.

Shonkwiler and Harris (1996) estimate a trivariate Poisson-gamma mixture model for the 1988 number of retail stores in three different sectors (Building materials and garden supply; Clothing; and Furniture) in each of 242 rural U.S. communities having populations between 100 and 5,000. The explanatory variables are the population size, the square root of the population size, per-capita income and the population density.

Finally, there are a number of applications related to the financial sector. Davutyan (1989) performs a time series analysis of the number of failed banks per year in the U.S. for 1947 - 1981, relating the bankruptcy risk to factors such as a measure of the absolute profitability of the economy, the relative profitability of the banking sector, as well as aggregate borrowing from the

Federal Reserve. Greene (1998) estimates a count data model using individual level data on the number of major derogatory reports in a sample of credits card holders. And Jaggia and Thosar (1993) study the determinants of the number of bids received by 126 U.S. firms that were targets of tender offers during the period 1978-1985 and whose management resisted takeover.

9.7 Labor Mobility

Labor mobility is a pervasive feature of market economies. Individuals typically hold several jobs during their working career. Topel and Ward (1992) report an average of 9 job changes during lifetime for male workers in the U.S. Own calculation for the German labor market, based on the *German Socio–Economic Panel*, indicate a distinctly lower average mobility of 3 male lifetime job changes. The sources of international differences in labor mobility are a research topic of substantial interest. A related question is why labor mobility differs so much between individuals within a country. And what can these differences tell us about the operation of the labor markets?

Explaining the variation in individual labor mobility has been a topic of interest in applied labor economics for more than 40 years. Early studies are Long and Bowyer (1953) and Silcock (1954). Recent studies include Börsch-Supan (1990), Topel and Ward (1992), Jung and Winkelmann (1993), and Winkelmann and Zimmermann (1993a, 1993b, 1994, 1998). The existing literature reports the following stable empirical findings.

1. Individual variation in mobility, as measured for instance by the variance of the number of job changes during a given period, is great. Hall (1982, p. 716) paraphrases this observation for the U.S.: "Though the U.S. labor market is justly notorious for high turnover (...) it also provides stable, near-lifetime employment to an important fraction of the labor force."
2. Most of the job changes occur at early stages of the career. In the US, an average of two out of three lifetime job changes occur during the first ten years after entering the job market (Topel and Ward, 1992). In Germany, about one out of two job changes falls within the first ten career years (own calculations using the German Socio-Economic Panel).
3. Labor mobility reduces with increasing tenure. Or, using Silcock's (1954, p. 430) words, "the amount of wastage decreases as the lenght of service increases".

Employment can be characterized in many ways: by occupation, employer, location, and position on the job ladder, to name but a few. Mobility in a broad sense is a change in any of these categories. Some types of mobility affect several categories at a time. For instance, a change of employer may require both a move to another city or region, and a change in occupation. On the other hand, geographic dislocation and moves on the job ladder may occur within a single firm. Regional mobility is studied in Börsch-Supan (1990).

Lazear (1990) addresses the issue of intra-firm job mobility. Here, as in Topel and Ward (1991) and Jung and Winkelmann (1993), labor mobility is defined as a change of employer. This event is referred to as a "job change".

9.7.1 Economics Models of Labor Mobility

Most analyzes of the determinants of individual labor mobility are in one way or another based on the *human capital theory* (See Becker, 1962, and Mincer 1962). The human capital theory states that workers invest in productivity enhancing skills as long as the cost is less then the present value of the expected future benefits. The return to human capital depends on the wage which, in competitive labor markets, equals marginal productivity (for a given type of human capital).

The human capital approach has been mainly used to model the dynamics of individual earnings over the life cycle. In particular, the theory implies a wage growth over the life cycle since initial earnings disadvantages during the time of human capital investments (which optimizing behavior places at early stages of the life cycle) are joined by higher wages during later periods. Further, if individuals are observed in a cross section, observed earnings differentials can be explained by variations in the stock of human capital, as measured by variations in the years of schooling, labor market experience, and tenure.

For the analysis of labor mobility, it is important to distinguish between two types of human capital. The first is *general* human capital, which is acquired through the education system. The second is *firm specific* human capital. It is acquired with current tenure and, as opposed to general human capital, it may not be transferable across employers. Thus, firm specific human capital creates a wedge between actual wages and potential outside wages. An increase in the wedge through larger firm specific investments reduces mobility.

The implications of human capital for labor mobility have been further explored by Jovanovic (1979a, 1979b) and Mincer and Jovanovic (1981). These authors emphasize the importance of imperfect information and heterogeneity. In particular, it is assumed that each worker has a nondegenerate productivity distribution across different firms or jobs. Human capital effects enter the model by determining the location (and possibly dispersion) of this distribution, and its shift over time. Further, the models are based on the following decision rule: a job change occurs if the expected present value of an alternative job is higher than the expected present value of the incumbent job (or if the difference exceeds transaction costs in case they exist).

A job change requires new information that changes the expectations of either the incumbent job or the outside offers. Two model types have been developed. In the first version, job changes occur as a result of new information about the current match (Jovanovic, 1979a). In this view, jobs are considered to be *experience* goods. The value of the match is unknown a priori but

reveals itself by experiencing the match. Thus, the experience provides new information which is for instance processed using a Bayesian updating rule. A job separation occurs if, compared to the initial evaluation, the present match is revealed to have a lower expected present value.

In the second type of models, job changes occur as a result of new information about the outside offers (Jovanovic, 1979b). In this view, jobs are considered to be *inspection* goods, i.e., the value of a match is known prior to the match. Here, job changes occur as new information about better paying outside jobs arises. The arrival rate of new information increases with the search effort which in turn increases search costs. Thus, an optimal search strategy can be established.

These models predict the following effects of standard human capital variables like education and labor market experience on labor mobility. The effect of education, as measured by the years of schooling, is ambiguous. First, as far as education creates general human capital, it should increase both inside and outside opportunities, i.e. (potential) wages, proportionally und thus leave mobility unaffected. Second, better general education creates skills that allow for a faster accumulation of firm specific human capital. Thus, for given tenure, individuals with higher education have a lower mobility. Third, in markets with imperfect information, better educated individuals should be better able to collect and process information. They tend to have a higher search efficiency and therefore lower transaction costs and higher mobility.

The models unambiguously predict an inverse relation between tenure and mobility. The negative correlation arises due to a wedge created by the accumulation of firm specific human capital over time or, if jobs are seen as experience goods, due to the operation of a sorting process.

Separating tenure and experience effects may be impossible. As Mincer and Jovanovic (1981) point out, a distinction has to be made between true experience effects and indirect effects via job tenure. Let the propensity to change job m be a function of both tenure ten and experience ex. Then

$$\frac{\mathrm{d}m}{\mathrm{d}ex} = \frac{\partial m}{\partial ten} \cdot \frac{\mathrm{d}ten}{\mathrm{d}ex} + \frac{\partial m}{\partial ex} \; . \tag{9.1}$$

Only $\partial m/\partial ex$ is a genuine experience effect. It is complemented by an indirect tenure effect since tenure grows with experience. Clearly, $0 < \mathrm{d}ten/\mathrm{d}ex < 1$, and mobility declines with experience also if there is no true experience effect but only a tenure effect, unless one controls for tenure. With count data this is generally not possible as job tenure (at the time of the job change) is not observed. Hence, the two effects are not separately identified. The reduced form effect of labor market experience picks up the combined effects of experience and tenure.

9.7.2 Previous Literature

Börsch-Supan (1990) studies the influence of education on labor and regional mobility using data from the PSID. Observations on 736 male individuals

are available for the period 1962-1982. Estimating Poisson regression models, Börsch-Supan finds that an increase in education reduces labor mobility, while it increases regional mobility. The reduction in labor mobility with increasing education is of considerable magnitude: The lowest education level has a predicted number of job changes that is about 50 percent higher than the predicted number of job changes for the highest education level. The conditional effect (i.e. after controlling for other characteristics in a multiple regression and evaluating the remaining variables at their sample means) is greater than the marginal effect obtained in a cross tabulation. Further, Börsch-Supan finds that experience has a negative effect on both types of mobility. As mentioned earlier, this finding might reflect tenure effects that cannot be controled for.

Merkle and Zimmermann (1992) use a German sample of labor force participants drawn from the unemployment register in 1977. The 1610 selected individuals answered questions on the number of employers and the previous number of unemployment spells during a five year period preceding the interview. The data are censored from above at five. Using Poisson and negative binomial regression models for censored data, Merkle and Zimmermann (1992) find that both the number of job changes and the number of unemployment spells increase with the education level, whereas these variables are affected in a concave way by previous labor market experience. Thus, their evidence is in conflict with the findings of Börsch-Supan. This apparent contradiction can be resolved when considering the differences in the sampling schemes. Sampling from the stock of unemployed as opposed to sampling from the labor force already tends to select less skilled individuals with a higher propensity to unstable labor relations. Within this group, better educated people might have higher re-employment chances, reducing their overall time spent in unemployment and increasing their turnover.

A further study of interest is Ebmer (1990) who looks at the determinants of offer arrival frequencies. The process of job mobility may be decomposed into two steps. In a first step, offers are made to the individual at a certain rate. In a second step, the individual decides whether or not to accept the offer. Usually, data on offer arrivals are not available. In Germany and Austria, however, job offers both for unemployed and employed individuals are mainly administered through a public placement service. Ebmer (1990) uses data on offers provided by the Austrian placement service, and, using Poisson and Negbin models, finds that the offer arrival rate falls with elapsed unemployment duration, which he interprets as discriminating behavior of labor exchange officials. Furthermore, his dataset allows to test for the assumption of Poisson arrival rates. This assumption is common in the search literature. The hypothesis is rejected although one cannot exclude that rejection is due to unobserved heterogeneity.

9.7.3 Data and Descriptive Statistics

The following sections illustrate the use of count data models for studying labor mobility in an empirical application using data from the *German Socio-Economic Panel* (GSOEP). Wagner, Burkhauser and Behringer (1993) provide a short introduction to the data set. The annual panel was first collected in 1984. The basic sampling units are households. The sample included 5921 households in 1984. Within each household, every person aged 16 or older is interviewed, resulting in 12,245 person records for 1984. The selection of households is stratified by nationality: One subsample consists of a random sample of the population living in Germany which is not of Turkish, Yugoslave, Greek, Italian or Spanish nationality. The proportion of non-Germans in this subsample of 9076 individuals is 1.5%. The second subsample of size 3169 includes 33% Turks, 18% Yugoslaves, 15% Greeks, 20% Italians and 13% Spaniards (Deutsches Institut für Wirtschaftsforschung 1990). All in all, the GSOEP oversamples the foreign population whose overall proportion was 7.5% in 1984 (Statistisches Bundesamt 1985).

The dependent variable is the number of employers and the number of unemployment spells during the ten year period 1974-84. This information is collected retrospectively in the first wave of the panel. In order to ensure that the analysis is based on persons with a reasonably strong labor force attachment, the sample is restricted to persons in employment in 1984 whose work career started before 1974. Women are excluded in order to minimize complications due to non-participation spells. Non-participation is known to be empirically relevant for women, and yet unobservable in the type of data studied here. Finally, self-employed persons and civil servants are excluded. The resulting sample has 1962 observations.

Using the information on the number of employers and the number of unemployment spells, two measures of labor mobility can be derived. First, assume that

i) people do not return to the same job (or employer) after a spell of unemployment, and
ii) individuals have been employed at the beginning of the period.

Then the number of employers minus the number of unemployment spells minus 1 measures the number of direct job-to-job transitions (without an intervening unemployment spell). Under the same assumptions, the number of indirect job changes (job-to-unemployment-to-new job transition) is simply equal to the total number of unemployment spells. A cross tabulation of **direct job changes** and **unemployment spells** is given in Table 9.1

There is a slight positive correlation between the two types of mobility ($\rho = 0.06$). For instance, the proportion of individuals having experienced at least one **direct job change** is greater for the group of individuals that did experience one **unemployment spell** than for the group that did experience no unemployment. The same holds true for **unemployment spells** vs.

Table 9.1. Frequency of Direct Changes and Unemployment

		Direct Job Changes											
		0	1	2	3	4	5	6	7	8	9	10	12	Total
U	0	1102	301	105	25	20	5	1	2	1		2		1564
n	1	146	79	21	10	1	3	2	1		1			264
e	2	34	16	6	6	2	2	1	1				1	69
m	3	20	4	1		2								27
p	4	7	2		2									11
l	5	6	2											8
o	6	2									1			3
y	7	2												2
m	8	3												3
e	9	3												3
n	10	7												7
t	15	1												1
	Total	1333	404	133	43	25	10	4	4	1	2	2	1	1962

direct job changes. For both **direct job changes** and **unemployment spells** the mode is at zero. The means are 0.54 and 0.37, respectively (See Table 9.2). The variance–mean relation is 2.16 for **direct job changes** and 3.32 for **unemployment spells**, indicating a tendency for overdispersion at the marginal level. This appears to provide a first check of the (non-)validity of the Poisson regression model, since conditional overdispersion violates the Poisson assumption. However, overdispersion at the marginal level is (theoretically) compatible with mean-variance equality conditional on covariates.

Sec. 9.7.1 defined the primary empirical question: What can individual characteristics tell us about individual propensities towards mobility, measured by the frequency of future **direct job changes** and **unemployment spells** ? The theoretical arguments developed in Chap. 9.7.1 suggested the main variables of interest: **Education** as measured by the years of schooling and previous professional **experience**. Further variables which have been used in the literature to control for individual heterogeneity in wages and mobility are occupational status, nationality, family status and union membership. The corresponding dummy variables are (Yes=1; Sample means in parentheses) **Qualified White Collar** (0.137), **Ordinary White Collar** (0.059), **Qualified Blue Collar** (0.501), **Ordinary Blue Collar** (0.304), **German** (0.668), **Single** (0.077), and **Union** (0.429). Exact definitions and measurement issues are given in the notes to Table 9.2.

Table 9.2 displays the 'gross' effect of these variables on the two types of labor mobility. The mean values in the 1^{st} and 3^{rd} column give the average number of **direct job changes** (**unemployment spells**) during the ten year period 1974–84 for the various classifications.

The most visible effect is certainly the strong reduction of mobility with increased labor market **experience**. Individuals at the beginning of their

career (less than 5 years of experience) have on average 3 times more **direct job changes**, and almost 2 times more **unemployment spells** over the next ten years, than individuals with more than 25 years of professional experience. Furthermore, the amount by which the mobility is reduced decreases with experience, i.e., there exists a convex pattern between experience and mobility.

Table 9.2. Mobility Rates by Exogenous Variables

	Direct Changes[1]		**Unemployment**[2]		
	Mean	Std.Dev.	Mean	Std.Dev.	Obs.[3]
by Occupational Status[4]					
Qualified White Collar	0.498	1.032	0.212	0.638	269
Ordinary White Collar	0.566	0.999	0.257	0.777	113
Qualified Blue Collar	0.540	1.120	0.431	1.285	983
Ordinary Blue Collar	0.553	1.069	0.377	1.036	597
by Nationality[5]					
German	0.466	0.974	0.367	1.194	1311
Foreign	0.688	1.270	0.390	0.928	651
by Family Status[6]					
Single	0.651	1.246	0.697	1.671	152
Married	0.530	1.071	0.348	1.049	1810
by Union Status[7]					
Union	0.440	0.964	0.273	0.898	841
Nounion	0.615	1.163	0.450	1.243	1121
by Professional Experience[8]					
-5 Years	0.954	1.478	0.578	1.361	372
6-15 Years	0.543	0.988	0.384	1.059	672
16-25 Years	0.407	0.965	0.259	0.935	659
26+ Years	0.274	0.735	0.338	1.195	259
by Educational Attainment[9]					
-10 Years	0.585	1.132	0.405	0.997	511
11-12 Years	0.514	1.027	0.447	1.351	876
13-18 Years	0.567	1.304	0.247	0.799	478
19+ Years	0.402	0.640	0.187	0.507	97
Total	0.539	1.080	0.372	1.112	1962

Source: *German Socio-economic Panel*, own calculations.
Notes:

1. **Direct Changes** give the number of direct job changes an individual has experienced during the period 1974-1984. A direct job change is defined by the number of employers minus the number of unemployment spells minus one. The information is obtained through a retrospective question.
2. **Unemployment** gives the number of unemployment spells an individual has experienced during the period 1974-1984. As *Direct Changes*, the information is obtained through a retrospective question.

3. Number of observations in the sample. The total sample size is 1962. The selection was conditional on being male, being part of the labor force during the period 1974-1984, and on being neither self-employed nor civil servant.
4. **Occupational Status** is measured upon entry into the labor market, i.e., it is the status in the first job.
5. The distribution of the **Nationality** reflects that the *German Socio-economic Panel* is a stratified panel: Foreigners are oversampled as compared to their share of the labor force in Germany. However, the sampling is exogenous and not choice based.
6. An individual is classified as **Single** if he is and always was a single, i.e. widowers and divorced are classified as married.
7. **Union** membership in 1985. Included are members of unions and comparable professional organizations.
8. Professional **experience** uses information on the year of entrance into the labor market, subtracting the latter from 1974, the start of the ten year period.
9. To obtain a continuous measure of the **Educational Attainment** the years of schooling are calculated using information on the various degrees obtained by an individual, and attributing to every degree a "typical" time it requires. For instance, a university degree takes on average 18 years of schooling. The years of schooling measure also includes the time spent in professional education, as long as it is a part of special training programs ("Lehre").

The effect of **education** on mobility is less uniform. Comparing individuals with less than 10 years of schooling and individuals with 13 to 18 years of schooling, there is almost no change in the average number of **direct job changes**. The number of **unemployment spells**, by contrast, is reduced by 40% for the more educated individuals.

Germans, union members, married individuals and qualified white collar workers have on average less **direct job changes** than foreigners, non-union members, singles and ordinary white collar or blue collar workers, respectively. The number of **unemployment spells** is higher for blue collar workers, singles, and non-union members than for white collar workers, married individuals, and union members, respectively. Nationality seems to have no effect on the frequency of unemployment.

Although the descriptive statistics provide some valuable information on the interaction between the variables, an interpretation in the light of the aforementioned theories is problematic. While the theoretical models establish *specific* effects, or effects that hold ceteris paribus, the descriptive statistics display the *gross* effects which mix specific contributions and contributions due to correlations with other explanatory variables. Thus, a multiple count data regression analysis is required in order to estimate the specific effect of a unit change in one explanatory variable on the expected number of job changes, holding everything else constant. Moreover, it allows to predict the mobility behavior for any given individual. Most importantly, though, it provides information on the underlying data generating process, i.e. the stochastic process governing mobility.

9.7.4 Regression Results

This chapter reports the results of various estimated models for the labor mobility data. For simplicity, we restrict our attention here to one of the two mobility measures, the number of direct job changes, from now on or simplicity referred to as the **number of job changes**. The models differ in the assumption on the underlying probability processes. To ensure comparability, the set of explanatory variables is kept identical in all cases. The explanatory variables include **education, experience, squared experience, union, single, German, qualified white collar, ordinary white collar,** and **qualified blue collar** worker.

The following models were estimated with **number of job changes** as dependent variable:

- Poisson
- Poisson-log-normal
- Negbin I, Negbin II and Geck$_k$
- robust Poisson
- Poisson-logistic
- hurdle Poisson
- probit-Poisson-log-normal
- finite mixture Poisson and finite mixture Negbin
- zero-inflated Poisson and zero-inflated Negbin

The full set of estimation results for the various models are listed in Tables D.1 – D.7 in Appendix D.

Poisson Results

The Poisson model is specified with a log-linear conditional expectation function. This means that the coefficients can be interpreted as semi-elasticities. Take the point estimate of -0.138, pertaining to the education effect, for illustrative purposes. Since the education variable is scaled (division by 10), we find that the estimated effect of 10 additional years of schooling is a reduction of the number of job changes by approximately 13.8 percent. The exact effect would be $[\exp(-0.138)-1]\times 100 = -12.9$, a 12.9 percent reduction. Similarly, based on the point estimate, each single additional year of education would reduce the number of job changes by 1.4 percent.

Sometimes, it is meaningful to compute absolute rather than relative marginal effects. This is in particular the case if one wants to compare marginal effects across models, where some of the models (such as the hurdle Poisson model or the zero-inflated Poisson model) may not have a log-linear conditional expectation function. We know from Chap. 3.1.4 that

$$\frac{d\,\mathrm{E}(y_i|x_i)}{d\,x_{ij}} = \exp(x_i'\beta)\beta_j$$

Thus, the marginal effect depends on the point in the covariate space where it is to be computed. It is common to take the sample mean, i.e., replace x_i by $\bar{x}$. We find for example that the marginal effect of education at the mean of the regressors is -0.067.

Of course, the point estimate of -0.138 and its associated relative or absolute marginal effects are subject to sampling variability. Indeed, one finds that **education** has no significant effect on **direct job changes**, since the t-ratio for the null hypothesis of no effect is about one, based on the Poisson standard errors estimated from inverting the Hessin matrix of the log-likelihood function (the computation of the standard error of the semi-elasticity and the marginal effect would need to be based on the delta rule).

Substantively, the result of 'no-effect' is compatible with the human capital view that education increases *general* human capital which in turn promotes outside and inside job opportunities alike. In other words, the level of education does not affect the probability of finding (being offered) a new job that is preferable to the current one. The finding is in contrast to Börsch-Supan (1990), who reports a negative and significant effect of the level of education on labor mobility. One possible explanation for the discrepancy is that he includes all job changes, also those with intervening spell of unemployment, whereas the results here are for the number of direct job-to-job transitions only.

The convex experience-mobility profile implied by the point estimates of the second order-experience polynomial is very plausible. It conforms to the stylized fact that job changes are much more likely to occur early in ones career. One year after entering the workforce, the expected job change rate has decreased 7.5 percent relative to the initial rate. The predicted job change rate further decreases with each additional year of experience, but at a decreasing rate. After 32 years of experience, the effect of experience on mobility reaches zero. If, instead of computing relative or percentage effects, one was interested in absolute changes, one would need to compute the marginal experience effect as follows: Let ex denote the variable 'experience', and $exsq$ denote the variable 'experience squared'. Then

$$\frac{\partial \mathrm{E}(y_i|x_i)}{\partial ex_i} = \exp(x_i'\beta)[\beta_{ex} + 2\beta_{exsq} ex_i] \tag{9.2}$$

This partial derivative depends on x_i. Evaluating (9.2) at the sample means, the marginal effect is given by -0.022. For an average individual, an additional year of experience decreases the expected number of job changes by 0.022.

Union membership reduces the expected number of job changes by 29 percent or, evaluating the effect as above, by 0.131 job changes during the ten year period, relative to non-unionized workers. This specific effect is smaller than the gross effect of 0.175 displayed in Table 9.2, reflecting the interactions between the variables. Finally, German nationality reduces the expected number of direct job changes, while the remaining variables have no significant effect on mobility.

While the previous remarks referred to the results of the Poisson regression, the findings display a remarkable robustness accross the various specifications. Table 9.3 compares the results for ten of the estimated models. The signs and the significance levels of the coefficients are mostly identical.

Table 9.3. Direct Job Changes: Comparison of Results

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Education	o	o	o	o	−−	o	−−	o	o	o
Experience	−−	−−	−−	−−	−−	−−	−−	o	−−	−−
Experience^2	++	++	++	++	+	++	o	o	o	o
Union	−−	−−	−−	−−	−	−−	−−		−−	o
Single	o	o	o	o	o	o		o	o	o
German	−−	−−	−−	−−	−	−−	−−		−−	−−
Qual. Wh.C.	o	o	o	o	o	o	o		o	o
Ord. Wh.C.	o	o	o	o	o	o	o		o	o
Qual. Bl.C.	+	o	o	o	o	o	o		o	+

Notes:

Dependent Variable: Direct Job Changes.
++ Positive sign and significant at the 5%-level.
+ Positive sign and significant at the 10%-level.
−− Negative sign and significant at the 10%-level.
− Negative sign and significant at the 5%-level.
o Insignificant.
(1) Poisson Regression.
(2) Robust Poisson Regression.
(3) Generalized Event Count Model (GEC_k)
(4) Hurdle Poisson (0/1).
(5) Hurdle Poisson (1+).
(6) Poisson lognormal.
(7) Poisson-Logistic Regression: Job Offers (Overlapping).
(8) Poisson-Logistic Regression: Acceptance.
(9) Two-components Poisson (Group 1)
(10) Two-components Poisson (Group 2)

What conclusions should we draw from this evidence? One might be tempted to argue that given the uniformity of the results and the different complexities of the models, the recommendation would be to choose the simplest model, in this case the standard Poisson regression. This approach would even find theoretical support by the consistency property of the Poisson regression as a PMLE. Therefore, the similarity of the findings in this application does not come too much as a surprise. However, this is only a part of the story. It neglects two important purposes of the econometric analysis: The use of the model for prediction and as a tool for learning about the underlying data generating process. We therefore proceed by investigating three further issues.

First, we assess which of the models has the best fit, a purely statistical exercise. Second, we investigate one possible reason for the superior performance of some models, applying the concept of marginal probability effects to show that a certain minimum flexibility in the distributional assumption is needed in order to account for the effect of the explanatory variables on the outcome distribution. And third, we return to a main theme of this book, namely that generalized models can be informative on interesting aspects of an underlying count mechanism, and that therefore structural inferences can be made. We will see to what extent this actually applies in the present context.

9.7.5 Model Performance

The Poisson regression model assumes that events (here: job changes) occur randomly over time, with a constant process intensity that is a deterministic function of individual covariates. The Poisson-log-normal, Negbin I, Negbin II and GEC_k models allow for unobserved heterogeneity. The remaining models relax the single-index structures in favor of a dual-index, or two-part structure. There are several ways of doing this. The hurdle models state that the intensity of the process switches conditional on the first occurrence. The Poisson-logistic model assumes a two-step process: In a first step, offers reach the individual according to a Poisson process. In a second step, the individual decides whether or not to accept the offer. The two-component models assume that the population consists of two latent groups, each one with its own regression function. The zero-inflated models use a logistic model to augment the probability of a zero relative to the base count data model.

The above models were estimated using an identical set of regressors and the following table lists the log likelihood, evaluated at the maximum likelihood parameter estimates.

Table 9.4. Number of Job Changes: Log Likelihood and SIC

	ℓ	K^1	SIC^2
Poisson	-2044.47	10	4164.76
Poisson-log-normal	-1866.80	11	3817.00
Negbin I	-1873.28	11	3829.96
Negbin II	-1878.63	11	3840.66
GEC_k	-1873.17	12	3837.32
Poisson-logistic (overlapping)[3]	-2039.35	13	4177.26
Poisson-logistic (non-overlapping)[3]	-2043.88	10	4163.58
Hurdle Poisson	-1928.00	20	4007.63
Probit Poisson-log-normal[3]	-1856.70	22	3880.20
two-components Poisson	-1868.16	21	3895.54
two-components Negbin II	-1856.05	23	3886.48
zero-inflated Poisson[3]	-1926.28	20	4004.19
zero-inflated Negbin II[3]	-1866.73	21	3892.68

Notes:

[1] K denotes the number of parameters in the model.
[2] Schwarz information criterion: $SIC = -2\ell + K \ln N$
[3] These models do not nest the Poisson model.

The log likelihood values can be used to formally test models against each other insofar as they are nested. For example, the Poisson model is nested in all unobserved heterogeneity-type models, as well as in the hurdle and the two-components Poisson models. From Tab. 9.4, the Poisson model is rejected by the various tests against any of the more general alternatives. This is clearly due to overdispersion. For instance, the GEC_k estimates a $\hat{sigma}^2$ of 0.892 with a standard error of 0.170. In the absence of over- or underdispersion, $\sigma^2 = 0$, but σ^2 is significantly greater than zero at any conventional significance level. The estimated k is not different from 0 either. However, it is significantly smaller than 1, providing evidence for the presence of a linear variance function as opposed to a quadratic one. Interestingly, the Poisson-lognormal model has a higher log-likelihood than either Negbin I or GEC_k. This suggests that the mixing distribution used to model unobserved heterogeneity is better described through a log-normal distribution than through a gamma distribution. Note, however, that the improved fit comes at the expense of increased computational complexity, since the integration requires numerical quadrature.

There are other nested model pairs in Tab. 9.4. For example, using likelihood ratio tests, the two-components Poisson model is rejected against the two-components Negbin II model (test statistic 24.22, p-value = 0.000); the non-overlapping Poisson-logistic model is rejected against the overlapping Poisson-logistic model (test statistic 9.06, p-value = 0.0285); and the zero-inflated Poisson model is rejected against the zero-inflated Negbin II model (test statistic 119.1, p-value = 0.000).

In other cases, Vuong's test for non-nested hypothesis can be used. For example, the Negbin I model and the Negbin II model are not nested. Since they both nest the standard Poisson model, they are overlapping rather than strictly non-nested, following the terminology of Vuong. Hence, a pre-test is required in order to establish that the two models are not equivalent. In this case, it is sufficient to show that the respective dispersion parameters are significantly different from zero. From Tab. D.2, we see that this is the case indeed. The null-hypothesis $H_0 : \sigma^2 = 0$ can be rejected in each model, using the asymptotic z test for instance. Next, the Vuong statistic proper can be computed. The formula was given in (3.89). The test statistic in this case is 0.999. It has a standard normal distribution, with the critical values being the usual $\alpha/2$ and $1 - \alpha/2$ quantiles. Hence, there is no evidence that the Negbin I model is significantly better than the Negbin II model. Note that this result differs from the conclusion based on the hyper model (GEC_k), where the Negbin II restriction could be rejected but the Negbin I restriction could not. The Vuong test has low power in finite samples. We also find that a test of

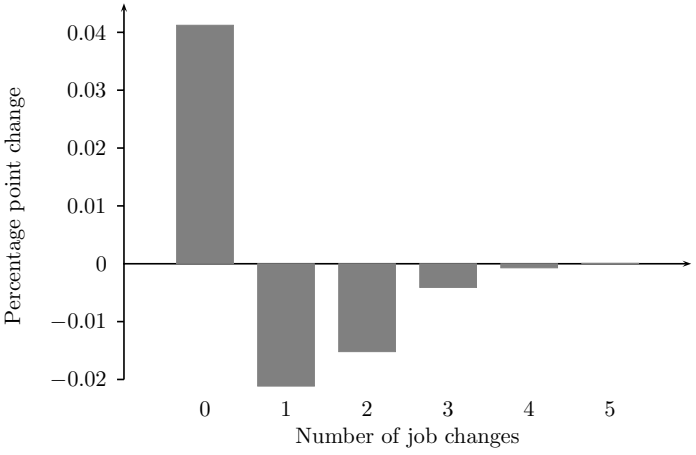
the Poisson-log-normal model against the Negbin I model is inconclusive. In this case, the Vuong test statistic is 1.255.

Finally, we can compare the models in Tab. 9.4 based on model selection criteria. The models have a different number of parameters, varying between a minimum of 10 and a maximum of 23. Using the Schwarz information criterion as a penalty function, it turns out the Poisson-log-normal model is the best model. The double index models lead to a large increase in the log-likelihood but this increase is more than offset by the larger number of additional parameters that needs to be estimated. In this application, it seems most important to use a model that allows for unobserved heterogeneity and overdispersion. Further generalizations are not dictated by the data. Nevertheless, as we will see next, these generalizations do offer some interesting insights into distributional effects of covariates and the underlying data generating process.

9.7.6 Marginal Probability Effects

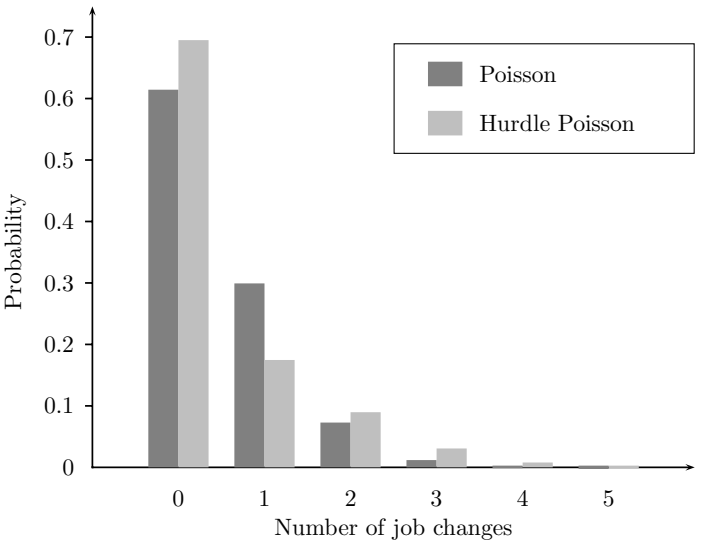
It was mentioned earlier that ten additional years of education are predicted to reduce the number of job changes by 0.067, based on the Poisson estimates (Tab. D.1) and the sample means of the explanatory variables. Fig. 9.1 shows how this mean effect arises. Ten additional years of education increase the probability of no job change by about 4 percentage points. The probability of one job change is reduced by 2 percentage points. The probability that a person reports two, three, four and so forth job changes is reduced at diminishing rates.

Fig. 9.1. Poisson Model: Marginal Probability Effect of a Unit Increase in Education



The Poisson marginal probability effects are very restrictive. For example, the sign of the effect can only change once from positive to negative, or vice versa. In order to see what would happen in a more flexible model, consider the hurdle Poisson model instead. The formula for computing the marginal mean effects in this double index model was given in (6.8). The formula for computing the marginal probability effects was given in (6.9). First, Fig. 9.2 shows the predicted probabilities of the Poisson and the hurdle Poisson model. As to be expected, the main difference is a larger probability of a zero in the hurdle model, corresponding to the phenomenon of unobserved heterogeneity/overdispersion/excess zeros in the data.

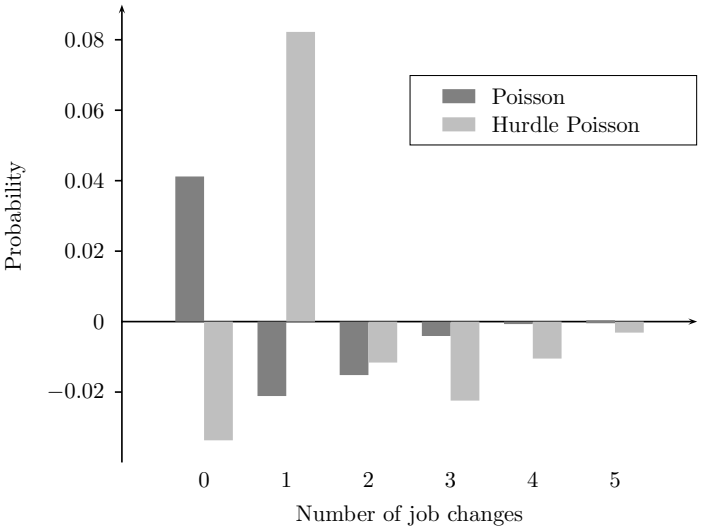
Fig. 9.2. Predicted Poisson and Hurdle Poisson Probabilities



But what can we say about the marginal effects? Consider the education effect as an example. Using (6.8), the estimated coefficients from Tab D.5, and evaluating the covariates at their sample means we obtain an effect of -0.071. This estimated mean effect is quite close to the effect in the Poisson model, -0.067. However, the same overall effect has very different distributional implications, as illustrated in Fig. 9.3. Based on the hurdle model, ten additional years of education reduce the probability of no job change by about 3 percentage points whereas the Poisson model predicts an increase. Similarly, in the hurdle Poisson model, we find that more education increases the probability of one job change, whereas the simple Poisson model predicts a decrease.

Using the hurdle model, we come thus to conclusions with regard to marginal probability effects that are diametrically opposite to those obtaines from the Poisson model. This is an illustration of the idea that an explanatory variable may have different marginal probability effects in different parts of the distribution (relative to a single index base model). If one ascribes substantive interest to these single outcomes, employing a sufficiently flexible model (that does not have the single crossing property) becomes imperative.

Fig. 9.3. Marginal Probability Effect of Education: Poisson and Hurdle Poisson



The hurdle model is only one among several possible generalizations. In the class of double index models, zero-inflated and two-components models would have similar advantages. From the perspective of allowing for maximal flexibility in the conditional probability distributions, there is no good reason to stop with double index models. The most general conceivable model would be a regression model based on a multinomial distribution. In this case, each outcome probability is parameterized as a separate function of the explanatory variables, subject to an adding-up constraint. For example, in the multinomial logit model,

$$p_{i1} = \frac{1}{1 + \sum_{k=2}^{J} \exp(x_i'\beta_k)}$$

$$p_{ij} = \frac{\exp(x_i'\beta_j)}{1 + \sum_{k=2}^{J} \exp(x_i'\beta_k)} \qquad j = 2, \ldots, J$$

where $j = 1, 2, \ldots, J$ are the J distinct counts observed in the sample. We immediately see two limitations of this model. First, it will only work if J is relatively moderate, since otherwise parameters will proliferate unduly. This problem could be mitigated in an ad-hoc way by grouping outcomes into classes. Second, the model does not allow the prediction of probabilities (or marginal probability effects) for outcomes that are not observed in the data. Relatedly, the model stands in no correspondance to an underlying count process. In short the multinomial logit model is not a count data model proper. Putting these reservations aside for a moment, one can use the model as a descriptive tool, obtaining the following result for the job change example.

Fig. 9.4. Marginal Probability Effect of Education: Hurdle Poisson and Multinomial Logit

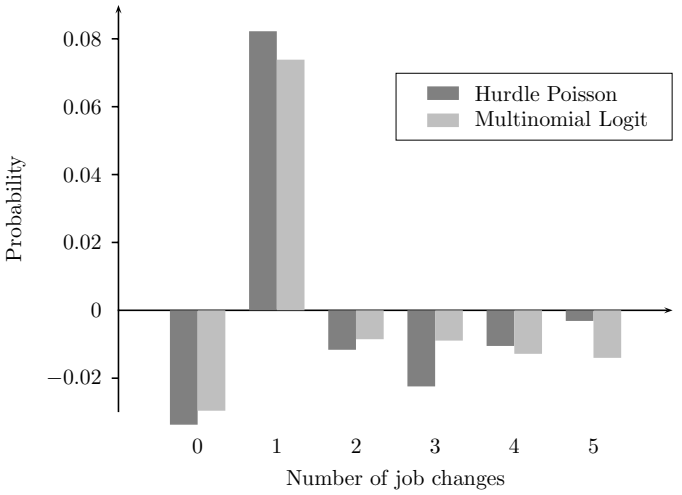
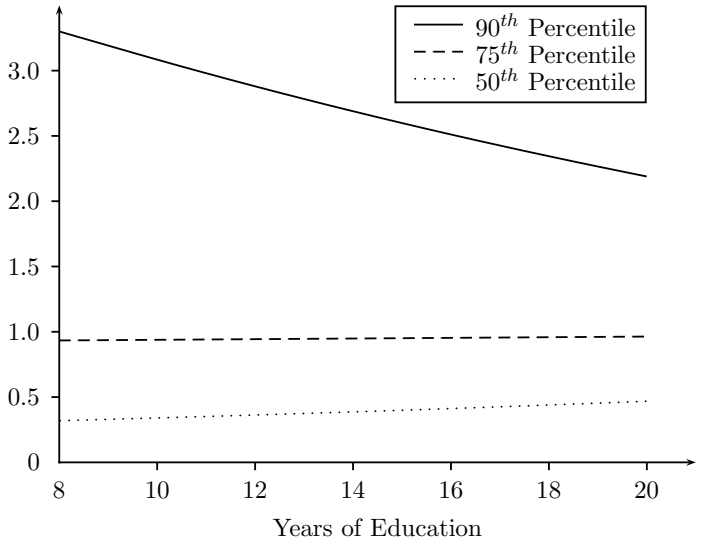


Fig. 9.4 compares the marginal probability effects of education in the hurdle model and in the multinomial logit model, everything as before evaluated at the sample means of the explanatory variables. Surprisingly, the discrepancy between the two response functions are minor. One can conclude that in this example, the double index hurdle model offers sufficient flexibility to capture how education changes the outcome distribution. More general models are not necessary.

An alternative way of capturing "non-standard" responses of whole distributions to changes in an explanatory variable is through quantile regression (Chap. 6.6). Consider the relative mean effect of education, a 13.8 percent decrease in the average number of joh changes per 10 additional years of ed-

ucation. If this effect was spread evenly over the distribution, and since the distribution is non-negative, one would expect to find a negative effect of additional education at all quantiles. Tab. D.8 shows the regression results for the 50-percent, the 75-percent and the 90-percent quantiles. We see that the effect of education is not so uniform. Quite to the contrary, the 50-percent quantile and the 75-percent quantile are positive functions of education. This is also seen from Fig. 9.5.

Fig. 9.5. 50/75/90 Percent Quantiles by Years of Education



This finding is indeed compatible with the hurdle and multinomial results above. More educated people are on average less mobile but they also have a larger probability of having changed job once, relative to less educated people.

9.7.7 Structural Inferences

Any empirical analysis can have one of two goals. A first goal is to describe the data as well as possible, estimate the effect of explanatory variables on certain key features, such as conditional mean or probability function, and to predict outcomes. A second goal is to learn something about the data generating process. This second, 'structural', approach requires stronger assumptions. Given the validity of these assumptions, one can then draw stronger inferences. In the following, we will discuss the idea behind such structural models in the

context of the analysis of labor mobility, using three examples: Poisson-logistic regression, zero-inflated Poisson models, and two-components Poisson models.

Poisson-Logistic Regression

As mentioned earlier, this model combines a latent process for counting event occurrences with an observation mechanism. A combination of the two then leads to the observed counts. In the context of labor mobility, one can reinterpret this set-up as one, where the latent model describes job offers, the reporting mechanism reflects whether a job offer is accepted or not, and the reported counts are equal to the number of accepted offers, i.e., the number of job changes, in a given period of time. Based on Tab. D.4, one may calculate the expected number of job offers for the sample. This number has to be greater than the number of accepted offers. For the overlapping specification, the predicted number of offers is 0.84 as opposed to 0.54 predicted job changes. Thus, a typical individual accepts roughly two out of three job offers.

Zero-Inflated Poisson Regression

A frequent substantive interpretation of the zero-inflated Poisson model is one where 'strategic' zeros are combined with 'incidental' ones. In the context of job search, one could think of this distinction as follows. Some individuals do not look for outside jobs at all, maybe because they are civil servants, or for some other reason. As a consequence, they are unlikely to get any outside offers. These are 'strategic' non-changers, because they have decided to keep their current job. In contrast, other individuals may be 'in the market'. These persons look for outside jobs. Some of them do not find an alternative job in a given period. These are then 'incidental' non-changers.

Within the structure of the zero inflated Poisson model (see Tab. D.7) one can compute the relative frequencies of the two types of workers. The model predicts model predicts 68.4 percent zeros, which is close to the 67.9 percent observed in the sample (this and the following predictions are computed first for each individual, given their covariates. Then, arithmetic means are calculated). The predicted proportion of strategic zeros is 50.9 percent. The probability of an incidental zero can be computed as $(1-0.509)\times 0.356 = 17.5$ percent. Hence, 74 percent of all zeros are strategic and 26 percent are incidental.

Two-Components Poisson Regression

The two-components Poisson model allows inferences to be drawn with respect to two subpopulations. From the results in Tab. D.6, we know that an estimated 93 percent of the population belong to Group 1, whereas 7 percent of the population belong to Group 2. One can compute the mean job change

rate for the two groups, based on sample means of the explanatory variables. Group 1 has a mean of 0.31 changes, whereas Group 2 has a mean of 2.8 changes. Thus, most individuals belong to the low-mobility group. One can furthermore study, how the response to explanatory variable differs between the two groups. Take the effect of education as an example. In the low mobility group ten additional years of education increase the number of job changes by a predicted 7.8 percent. In the high mobility group ten additional years of education reduce the number of job changes by a predicted 36.8 percent. These differential effects are compatible with the results in the previous chapter, for example based on quantile regression, where a large negative effect of education was found at the 90th percentile, and a small positive effect was found at the median.

A

Probability Generating Functions

This appendix is based on Feller (1968, Chap. XI and Chap. XII). Let X be a random variable taking values $j \in I\!N_0$ with $P(X = j) = p_j$. Upper case letters X, Y, and Z denote a random variable, while lower case letters j and k denote a realization. $p_{j \in I\!N_0}$ is called the *probability function*, while $F_{i \in I\!N_0} = P(X \leq i)$ is called the *distribution function*.

Definition 1. .
Let X be a random variable defined over the non-negative integers. The probability generating function (PGF) is given by the polynomial

$$\mathcal{P}^{(X)}(s) = p_0 + p_1 s + p_2 s^2 + \ldots = \sum_{j=0}^{\infty} p_j s^j = \mathrm{E}(s^X) \tag{A.1}$$

The function $\mathcal{P}(s)$ is defined by the $p'_j s$ and, in turn, defines the $p'_j s$ since a polynomial expansion is unique.

Example: Let X have a binomial distribution function with parameters n and p, $p_j = 0$ for $j > n$ (writing $X \sim B(n, p)$). The probability generating function is given by

$$\mathcal{P}(s) = \sum_{j=0}^{n} \binom{n}{j} (ps)^j q^{n-j} = (q + ps)^n \tag{A.2}$$

If it is not clear out of the context which random variable is meant, we write $\mathcal{P}^{(X)}$ where X is the random variable. An important property of a PGF is that it converges for $|s| \leq 1$ since $\mathcal{P}(1) = \sum_{j=0}^{\infty} p_j = 1$. The PGF can be used to directly derive the probability function of the random variable, as well as its moments. Single probabilities can be calculated as

$$P(X = j) = p_j = (j!)^{-1} \left. \frac{\mathrm{d}^j \mathcal{P}}{\mathrm{d}s^j} \right|_{s=0} \tag{A.3}$$

Example: A binomial distributed random variable has PGF $\mathcal{P}(s) = (q+ps)^n$. Thus,

$$\begin{aligned} P(X=0) &= \mathcal{P}(0) = q^n \\ P(X=1) &= \mathcal{P}'(0) = nq^{n-1}p^1 \\ P(X=2) &= (2!)^{-1}\mathcal{P}''(0) = (2!)^{-1}n(n-1)q^{n-2}p^2 \\ &\vdots \qquad \vdots \end{aligned}$$

The expectation $\mathrm{E}(X)$ satisfies the relation

$$\mathrm{E}(X) = \sum_{j=0}^{\infty} jp_j = \mathcal{P}'(1) \tag{A.4}$$

Example: A binomial distributed random variable has mean

$$\begin{aligned} \mathcal{P}'(1) &= np(q+p)^{n-1} \\ &= np \end{aligned}$$

Calculating first

$$\mathrm{E}[X(X-1)] = \sum_{j=1}^{\infty} j(j-1)p_j = \mathcal{P}''(1) \tag{A.5}$$

the variance is obtained as

$$\begin{aligned} \mathrm{Var}(X) &= \mathrm{E}[X(X-1)] + \mathrm{E}(X) - [\mathrm{E}(X)]^2 \\ &= \mathcal{P}''(1) + \mathcal{P}'(1) - [\mathcal{P}'(1)]^2 \end{aligned} \tag{A.6}$$

Example: A binomial distributed random variable has variance

$$\begin{aligned} \mathrm{Var}(X) &= n(n-1)p^2 + np - (np)^2 \\ &= np(1-p) \end{aligned}$$

Proposition 1. *Let X be a random variable defined over the non-negative integers with probability distribution $P(X=j) = p_j$, $j = 0, 1, \ldots$. Let X_T be a positive random variable with truncated-at-zero probability distribution $P(X_T = j) = p_j/(1-p_0)$, $j = 1, 2, \ldots$. The probability generating function of the truncated-at-zero distribution of X_T is given by*

$$\mathcal{P}_T(s) = \frac{\mathcal{P}(s) - \mathcal{P}(0)}{1 - \mathcal{P}(0)} \tag{A.7}$$

Proof: (A.7) follows directly from the definition of the probability generating function:

$$\mathcal{P}_T(s) = \mathrm{E}(s^{X_T}) = \sum_{j=1}^{\infty} \frac{p_j}{1-p_0} s^j$$

where $p_0 = \mathcal{P}(0)$.

There exists a close relationship between the probability generating function and the moment generating function $\mathcal{M}(t)$:

$$\mathcal{M}(t) = \mathrm{E}(e^{tX}) = \mathcal{P}(e^t) \tag{A.8}$$

While the moment generating function is a concept that can be used for any distribution with existing moments, the probability generating function is defined for non-negative integers. Since $s = e^t = 1$ if and only if $t = 0$, we obtain $\mathrm{E}(X) = \mathcal{P}'(1) = \mathcal{M}'(0)$.

In the same way as in (A.1) one can define a *bivariate probability generating function*.

Definition 2. .
Let X, Y be a pair of integer-valued random variables with joint distribution $P(X = j, Y = k) = p_{jk}\,, j, k \in \mathbb{N}_0$. The bivariate probability generating function is given by:

$$\mathcal{P}(s_1, s_2) = \sum_{j=0}^{\infty} \sum_{k=0}^{\infty} p_{jk} s_1^j s_2^k = \mathrm{E}(s_1^X s_2^Y) \tag{A.9}$$

Proposition 2. *The probability generating functions of the marginal distributions $P(X = j)$ and $P(Y = k)$ are $\mathcal{P}(s, 1) = E(s^X)$ and $\mathcal{P}(1, s) = E(s^Y)$, respectively.*

Proposition 3. *The probability generating function of $X + Y$ is given by $\mathcal{P}(s, s) = E(s^{X+Y})$.*

Proposition 4. *The variables X and Y are independent if and only if $\mathcal{P}(s_1, s_2) = \mathcal{P}(s_1, 1)\mathcal{P}(1, s_2)$ for all s_1, s_2.*

Probability generating functions can be used to establish the distribution of a sum of independent variables. This is also called a *convolution*. Using **Proposition 3** and **Proposition 4**, the probability generating function of $Z = X + Y$ is given by:

$$\mathcal{P}^{(Z)}(s) = \mathrm{E}(s^Z) = \mathrm{E}(s^{X+Y}) = \mathrm{E}(s^X s^Y) \overset{(\star)}{=} \mathrm{E}(s^X)\mathrm{E}(s^Y) \tag{A.10}$$

where $(\star)$ follows from the independence assumption.

Example: Let X have a binomial distribution function with $B(1, p)$. Consider the convolution $Z = \underbrace{X + \ldots + X}_{n-\text{times}}$. Then:

$$\mathcal{P}^{(Z)}(s) = (q + ps)^n \tag{A.11}$$

Z has a binomial distribution function $B(n, p)$. Conversely, the binomial distribution is obtained by a convolution of identically and independently distributed Bernoulli variables.

B

Gauss-Hermite Quadrature

This appendix describes the basic steps required for a numerical evaluation of the likelihood function of count data models with unobserved heterogeneity of the log-normal type. The method is illustrated for the Poisson-log-normal model, although a similar algorithm can be used to estimate the models with endogenous selectivity presented in Chap. 5.2. Butler and Moffitt (1982) discuss Gauss-Hermite quadrature in the context of a panel probit models. Million (1998) points out that the Poisson-log-normal integral can be approximated using Gauss-Laguerre and Gauss-Legendre polynomials as well, and he evaluates the relative performance of the three methods. Crouch and Spiegelman (1990) discuss numerical integration in the related logistic-normal model.

Starting point for Gauss-Hermite quadrature is the integral

$$\int_{-\infty}^{\infty} f(y|x,\beta,\varepsilon)g(\varepsilon|\sigma^2)\mathrm{d}\varepsilon \tag{B.1}$$

that cannot be solved by analytical methods. However, assume that by appropriate change of variable, B.1 can be brought into the form

$$\int_{-\infty}^{\infty} h(\nu;y,x,\beta,\sigma^2)\exp(-\nu^2)\mathrm{d}\nu \tag{B.2}$$

In this case, Gauss-Hermite quadrature can be applied to numerically evaluate the integral (B.1), and thus the marginal likelihood $L(y|x)$. Once the evaluation has been done, the logarithm $\ln L(y|x)$ can be passed on to a maximizer that uses numerical derivatives in order to find the maximum likelihood estimators $\hat{\beta}$ and $\hat{\sigma}^2$.

The Poisson-log-normal model has the following components (see also Chap. 4.2):

$$f(y|\varepsilon) = \frac{\exp(-\exp(x'\beta+\varepsilon))\exp(x'\beta+\varepsilon)^y}{y!}$$

where $\varepsilon \sim N(0,\sigma^2)$, i.e.,

$$f(\varepsilon) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{\varepsilon}{\sigma}\right)^2}$$

Change of variable from ε to ν where

$$\nu = \frac{\varepsilon}{\sqrt{2}\sigma}$$

has inverse $\varepsilon = \nu\sqrt{2}\sigma$ and Jacobian $\mathrm{d}f(\nu)/\mathrm{d}\nu = \sqrt{2}\sigma$. Therefore

$$g(\nu) = \frac{1}{\sqrt{\pi}} e^{-\nu^2}$$

and

$$f(y|\nu)g(\nu) = \frac{\exp(-\exp(x'\beta + \nu\sqrt{2}\sigma))\exp(x'\beta + \nu\sqrt{2}\sigma)^y}{\sqrt{\pi}y!} e^{-\nu^2}$$

Let

$$h_i(\nu) = \frac{\exp(-\exp(x_i'\beta + \nu\sqrt{2}\sigma))\exp(x_i'\beta + \nu\sqrt{2}\sigma)^{y_i}}{\sqrt{\pi}y_i!}$$

where the subscript i reminds us that this function depends on observations y_i and x_i. Then the Gauss-Hermite approximation to the integral $B.1$ is obtained as

$$\begin{aligned} L_i^{gh} &= \int_{-\infty}^{\infty} h_i(\nu)\exp(-\nu^2)\mathrm{d}\nu \\ &\approx \sum_{j=1}^{n} w_j h_i(\nu_j) \end{aligned}$$

where w_j are weights and ν_j are the evaluation points. The likelihood function for n independent observations is given by

$$L^{gh} = \prod_{i=1}^{n}\sum_{j=1}^{n} w_j h_i(\nu_j)$$

Weight factors and abscissas for 20-point quadrature are given in Tab. B.1 (Source: Abramowitz and Stegun, 1964, p. 924).

Table B.1. Abcissas and Weight Factors for 20-point Gauss-Hermite Integration

u_i	w_i
-5.3874809	2.2939000e-13
-4.6036824	4.3993400e-10
-3.9447640	1.0860000e-07
-3.3478546	7.8025500e-06
-2.7888061	0.00022833863
-2.2549740	0.0033243773
-1.7385377	0.024810521
-1.2340762	0.10901721
-0.73747373	0.28667551
-0.24534071	0.46224367
0.24534071	0.46224367
0.73747373	0.28667551
1.2340762	0.10901721
1.7385377	0.024810521
2.2549740	0.0033243773
2.7888061	0.00022833863
3.3478546	7.8025500e-06
3.9447640	1.0860000e-07
4.6036824	4.3993400e-10
5.3874809	2.2939000e-13

Source: Abramowitz and Stegun, 1964, p. 924

C

Software

Most statistical and econometric software distributions contain built-in procedures for standard count data models, such as the Poisson and the negative binomial regression models. Development in the software sector is fast, and specific recommendations risk to become outdated very quickly. Nevertheless, there are a few general points that should be of help to anyone interested in working with count data and estimating the models presented in this book.

Within the econometrics research community, GAUSS traditionally has been the major development tool. GAUSS is mostly a programming environment, but specialised procedures are available both as part of the general distribution, and through web sites and mailing lists. For example, the "count" module allows the estimation of seemingly unrelated regression models, of various types of negative binomial models as well as hurdle Poisson models. Yet, the development of this module has stalled for some time, and the latest models are not available.

Two alternative programs with a much more ambitious offering in this area are STATA and LIMDEP. This appendix is not intended as a comprehensive review of available software for count data, and there may be other software with similar or even broader scope. Yet, the possibilities that these two packages offers should be closely scrutinized by anyone seriously interested in count data applications who wants to apply up-to-date methods without doing the programming for herself. In fact, most of the models discussed in this book are easily estimated with STATA or LIMDEP, providing little support for those who resort to the most basic models in want of available software for the more appropriate ones.

The following short summary refers to STATA release 7.0. This release includes built-in procedures, apart from the standard Poisson and Negbin models (in its various parameterizations, as Negbin I, Negbin II or with more flexible variance function), for zero-inflated Poisson and zero-inflated negative binomial models, and for fixed and random effects panel count data models. Random effects models include the negative binomial panel model (with fixed or random effects) but also the panel Poisson-log-normal model. This proce-

dure can also be used in cross sections to estimate the standard Poisson-log-normal model that frequently has a better fit than the Negbin model. Hurdle Poisson or negative binomial models are not included in the standard distribution. However, they can be estimated using routines on truncated-at-zero models authored by Joseph Hilbe and described in the *Stata Technical Bulletin* Nr. 47. Most procedures include options for the computations of robust standard errors (to perform pseudo maximum likelihood estimation) as well as account for clustered sampling.

The latest version of LIMDEP is release 8.0. Apart from the standard count data models, its capabilities include the estimation of sample selection models by maximum likelihood, parametric models for underreporting where the observed counts represent only the reported fraction of the total events which have occurred, and maximum likelihood estimation of various types of hurdle models and zero-inflated models. LIMDEP and STATA are both quite versatile in the area of count data modelling.

D

Tables

Table D.1. Number of Job Changes: Poisson and Poisson-Log-Normal

	Poisson	Poisson-log-normal	Mean
Constant	0.501**	0.072	1
	(0.158)	(0.227)	
Education$*10^{-1}$	-0.138	-0.120	1.216
	(0.137)	(0.187)	
Experience$*10^{-1}$	-0.770**	-0.846**	1.460
	(0.111)	(0.155)	
Experience$^2 * 10^{-2}$	0.119**	0.127*	2.943
	(0.037)	(0.050)	
Union	-0.292**	-0.324**	0.429
	(0.065)	(0.088)	
Single	-0.050	-0.093	0.077
	(0.108)	(0.153)	
German	-0.368**	-0.390**	0.668
	(0.076)	(0.104)	
Qualified White Collar	0.067	-0.002	0.137
	(0.131)	(0.179)	
Ordinary White Collar	0.185	0.190	0.058
	(0.147)	(0.207)	
Qualified Blue Collar	0.147	0.112	0.501
	(0.082)	(0.114)	
σ^2		1.048**	
		(0.048)	
Log likelihood	-2044.47	-1866.80	
Log likelihood ($\beta_1, \ldots, \beta_9 = 0$)	-2155.40	-1934.53	
Number of Observations	1962		

Source: *German Socio-Economic Panel*, Wave A/1984; own calculations.
Note: Asymptotic standard errors in parentheses.

Table D.2. Number of Job Changes: Negative Binomial Models

	Negbin I	Negbin II	GEC_k
Constant	0.341	0.616**	0.380
	(0.191)	(0.224)	(0.212)
Education$*10^{-1}$	0.008	-0.179	-0.011
	(0.162)	(0.187)	(0.180)
Experience$*10^{-1}$	-0.762**	-0.786**	-0.775**
	(0.139)	(0.152)	(0.144)
Experience$^2 * 10^{-2}$	0.113*	0.118*	0.115*
	(0.046)	(0.048)	(0.047)
Union	-0.274**	-0.308**	-0.283**
	(0.080)	(0.087)	(0.084)
Single	-0.114	-0.054	-0.108
	(0.139)	(0.152)	(0.141)
German	-0.316**	-0.404**	-0.331**
	(0.097)	(0.102)	(0.103)
Qualified White Collar	-0.022	0.043	-0.013
	(0.163)	(0.174)	(0.173)
Ordinary White Collar	0.213	0.188	0.214
	(0.176)	(0.201)	(0.181)
Qualified Blue Collar	0.086	0.132	0.094
	(0.103)	(0.111)	(0.107)
σ^2	0.823**	1.378**	0.892**
	(0.088)	(0.137)	(0.080)
k			0.139
			(0.281)
Log likelihood	-1873.28	-1878.63	-1873.17
Number of Observations	1962		

Source: *German Socio-Economic Panel*, Wave A/1984; own calculations.
Notes: Asymptotic standard errors in parentheses. For $\sigma^2 > 0$ and $k = 0$, the GEC_k model coincides with the Negbin I model. For $\sigma^2 > 0$ and $k = 1$, the GEC_k model coincides with the Negbin II model.

Table D.3. Number of Job Changes: Robust Poisson Regression

	Coefficient	t_{Poisson}	Robust t-Values t_{WHITE}	t_{LVF}	t_{QVF}
Constant	0.501	3.167	2.617	2.229	2.304
Education$*10^{-1}$	-0.138	-1.006	-0.823	-0.707	-0.749
Experience$*10^{-1}$	-0.770	-6.929	-4.830	-4.877	-5.055
Experience$^2 * 10^{-2}$	0.119	3.269	2.385	2.301	2.486
Union	-0.292	-4.499	-3.115	-3.167	-3.385
Single	-0.050	-0.460	-0.309	-0.323	-0.326
German	-0.368	-4.843	-2.892	-3.409	-3.503
Qualified White Collar	0.067	0.514	0.343	0.361	0.384
Ordinary White Collar	0.185	1.255	0.964	0.883	0.917
Qualified Blue Collar	0.147	1.794	1.261	1.263	1.308
Log likelihood	-2044.47				
Number of Observations	1962				

Notes:

Three alternative methods to calculate robust standard errors (and thus robust t-values) were given in Chap. 3.3.3. t_{LVF} and t_{QVF} are based on the assumption of a quadratic and linear variance function, respectively, while the White method makes no explicit assumption.

Table D.4. Number of Job Changes: Poisson-Logistic Regression

	a) Overlapping		**b) Non Overlapping**	
Variable	Offers	Acceptance	Offers	Acceptance
Constant	0.812		1.151	
	(3.746)		(9.740)	
Education$*10^{-1}$	-0.322	3.732		-0.260
	-2.073)	(1.582)		(-1.633)
Experience$*10^{-1}$	-0.668	-6.044		-1.068
	(-4.804)	(-1.221)		(-7.678)
Experience$^2 * 10^{-2}$	0.071	3.321		0.175
	(1.382)	(1.132)		(3.920)
Union	-0.291		-0.290	
	(-4.477)		(-4.470)	
Single		0.379		-0.068
		(0.153)		(-0.460)
German	-0.397		-0.355	
	(-5.112)		(-4.708)	
Qualified White Collar	0.069		0.088	
	(0.452)		(0.684)	
Ordinary White Collar	0.178		0.195	
	(1.125)		(1.328)	
Qualified Blue Collar	0.132		0.156	
	(1.389)		(1.919)	
Log likelihood	-2039.35		-2043.88	
Observations	1962			

Notes:

Asymptotic t-values in parentheses.

Table D.5. Number of Job Changes: Hurdle Count Data Models

	Hurdle Poisson		Probit-Poisson-log-normal	
Variable	1+/0	1+	1+/0	1+
Constant	-0.069	1.163	0.269	0.799
	(0.202)	(0.245)	(0.157)	(0.666)
Education$*10^{-1}$	0.133	-0.600**	0.094	-0.764**
	(0.170)	(0.218)	(0.128)	(0.324)
Experience$*10^{-1}$	-0.758**	-0.403**	-0.629**	-0.544
	(0.148)	(0.156)	(0.111)	(0.405)
Experience$^2 * 10^{-2}$	0.107**	0.085	0.098**	0.103
	(0.048)	(0.050)	(0.034)	(0.088)
Union	-0.268**	-0.167*	-0.205**	-0.230
	(0.084)	(0.097)	(0.061)	(0.189)
Single	-0.194	0.192	-0.149	0.195
	(0.149)	(0.147)	(0.114)	(0.249)
German	-0.330**	-0.206**	-0.254**	-0.223
	(0.101)	(0.108)	(0.076)	(0.208)
Qualified White Collar	-0.071	0.271	-0.076	0.283
	(0.170)	(0.196)	(0.125)	(0.285)
Ordinary White Collar	0.239	-0.039	0.200	-0.057
	(0.185)	(0.236)	(0.143)	(0.336)
Qualified Blue Collar	0.069	0.184	0.042	0.167
	(0.109)	(0.117)	(0.081)	(0.172)
σ^2			0.932**	
			(0.156)	
ρ			0.212	
			(0.893)	
Log-likelihood	-1928.00		-1856.70	
Observations	1962			

Notes:

Asymptotic standard errors in parentheses.
Hurdle negbin results are nor displayed because of convergence problems.

Table D.6. Number of Job Changes: Finite Mixture Models

	2-components Poisson		2-components Negbin II	
Variable	group 1	group 2	group 1	group 2
Constant	-0.000	2.229**	1.047*	0.154
	(0.226)	(0.433)	(0.630)	(0.458)
Education$*10^{-1}$	0.078	-0.368	-0.648	0.243
	(0.184)	(0.374)	(0.425)	(0.307)
Experience$*10^{-1}$	-0.857**	-0.541**	-0.371	-1.140**
	(0.172)	(0.231)	(0.346)	(0.296)
Experience$^2 * 10^{-2}$	0.104*	0.081	0.050	0.138
	(0.060)	(0.074)	(0.099)	(0.105)
Union	-0.309**	-0.207	-0.259	-0.328**
	(0.095)	(0.141)	(0.181)	(0.152)
Single	-0.156	0.057	0.200	-0.274
	(0.158)	(0.229)	(0.325)	(0.267)
German	-0.351**	-0.478**	-0.609**	-0.101
	(0.114)	(0.158)	(0.236)	(0.229)
Qualified White Collar	-0.037	0.168	0.320	-0.263
	(0.192)	(0.288)	(0.386)	(0.327)
Ordinary White Collar	0.253	0.103	-1.037	0.609*
	(0.202)	(0.358)	(0.945)	(0.353)
Qualified Blue Collar	0.082	0.317*	0.393	-0.130
	(0.124)	(0.173)	(0.263)	(0.224)
σ^2			2.096**	0.146
			(0.949)	(0.281)
π_1	0.930**		0.395**	
	(0.013)		(0.158)	
Log-likelihood	-1868.16		-1856.05	
Observations	1962			

Notes:

Asymptotic standard errors in parentheses.
Hurdle Negbin I results are nor displayed because of convergence problems.

Table D.7. Number of Job Changes: Zero Inflated Count Data Models

	zero-inflated Poisson		zero-inflated Negbin II	
Variable	logit	Poisson	logit	Negbin II
Constant	1.132	-0.303	0.483*	-7.390**
	(0.245)	(0.529)	(0.255)	(2.777)
Education$*10^{-1}$	-0.583**	1.016**	-0.262	-0.746
	(0.203)	(0.455)	(0.216)	(1.152)
Experience$*10^{-1}$	-0.373**	1.035**	-0.613**	4.535**
	(0.153)	(0.312)	(0.192)	(1.715)
Experience$^2 * 10^{-2}$	0.072	-0.157	0.129**	-.759**
	(0.049)	(0.091)	(0.062)	(0.346)
Union	-0.158	0.293	-0.253**	0.351
	(0.097)	(0.179)	(0.102)	(0.465)
Single	0.151	0.461	0.066	1.368
	(0.154)	(0.297)	(0.169)	(0.962)
German	-0.173	0.435**	-0.236**	1.306
	(0.106)	(0.211)	(0.118)	(0.788)
Qualified White Collar	0.272	0.519	0.178	1.228
	(0.189)	(0.354)	(0.206)	(0.843)
Ordinary White Collar	-0.123	-0.870	0.025	11.967
	(0.243)	(0.711)	(0.203)	(349.869)
Qualified Blue Collar	0.166	0.094	0.151	0.193
	(0.115)	(0.219)	(0.129)	(0.614)
σ^2			1.103	
			(0.146)	
Log-likelihood	-1926.28		-1866.73	
Observations	1962			

Notes:

Asymptotic standard errors in parentheses.

Table D.8. Number of Job Changes: Quantile Regressions

	$Q_z(0.5, x)$	$Q_z(0.75, x)$	$Q_z(0.9, x)$
Constant	-0.181	1.138	1.768
	(0.468)	(0.475)	(0.272)
Education$*10^{-1}$	0.319	0.026	-0.343
	(0.373)	(0.343)	(0.207)
Experience$*10^{-1}$	-1.346	-1.413	-0.721
	(0.249)	(0.258)	(0.196)
Experience$^2 * 10^{-2}$	0.288	0.220	0.066
	(0.069)	(0.083)	(0.054)
Union	-0.388	-0.395	-0.336
	(0.193)	(0.187)	(0.117)
Single	-0.469	-0.191	-0.128
	(0.324)	(0.248)	(0.220)
German	-0.479	-0.522	-0.209
	(0.246)	(0.213)	(0.162)
Qualified White Collar	-0.144	-0.063	-0.020
	(0.304)	(0.302)	(0.237)
Ordinary White Collar	0.240	0.312	-0.061
	(0.319)	(0.340)	(0.188)
Qualified Blue Collar	-0.142	-0.078	-0.082
	(0.212)	(0.179)	(0.137)
Observations	1962		

Notes:

Bootstrap standard errors in parentheses (50 replications).

References

Abramowitz, M. and I. A. Stegun 1964, *Handbook of Mathematical Functions*, National Bureau of Standards, Applied Mathematics Series Nr. 55, Washington, D.C.

Aitchison, J. and C.H. Ho 1989, "The multivariate Poisson-log normal distribution," *Biometrika* 76(4): 643-653.

Aitkin, M 1999, "A general maximum likelihood analysis of variance components in generalized linear models," *Biometrics* 55, 117-128.

Albert, J.H. and P.A. Pepple 1989, "A Bayesian approach to some overdispersion models," *Canadian Journal of Statistics* 17: 333-344.

Alfo, M. and G. Trovato 2004, "Semiparametric mixture models for multivariate count data, with application," *Econometrics Journal* 7: 426-454.

Allison, P.D. 1984, *Event History Analysis: Regression for Longitudinal Event Data*, University Paper No. 46, Sage Publications: Beverly Hills.

Al-Osh, M.A. and A.A. Alzaid 1987, "First order integer valued autoregressive (INAR(1)) process," *Journal of Time Series Analysis* 8: 261-275.

Al-Osh, M.A. and A.A. Alzaid 1988, "Integer-valued moving average (INMA) process," *Statistical Papers* 29: 281-300.

Al-Qudsi, S. 1998a, "The demand for children in Arab countries: Evidence from panel and count data models," *Journal of Population Economics* 11(3): 435-452.

Al-Qudsi, S. 1998b, "Labour participation of Arab women: Estimates of the fertility to labour supply link," *Applied Economics* 30: 931-941.

Alvarez, B. and M.A. Delgado 2002, "Goodness-of-fit techniques for count data models: an application to the demand for dental care in Spain, " *Empirical Economics* 27: 543-567.

Amemiya, T. 1985, *Advanced Econometrics*, Harvard University Press, Cambridge.

Andrews, D.W.K. 1988, "Chi-squared diagnostic tests for econometric models: introduction and applications," *Journal of Econometrics* 37: 135-156.

Angrist, J.D. 2001, "Estimation of limited dependent variable models with dummy endogenous regressors: simple strategies for empirical practice," *Journal of Business and Economic Statistics* 19: 2-28.

Aptech 1994, *GAUSS Command Reference Manual*, Aptech: Maple Valley, WA.

Arulampalam, W. and A.L. Booth 1997, "Who gets over the training hurdle? a study of the training experiences of young men and women in Britain," *Journal of Population Economics* 10(2): 197-217.

Atella, V. and F.C. Rosati 2000, "Uncertainty about children's survival and fertility: A test using Indian microdata," *Journal of Population Economics* 13(2): 263-278.

Bago d'Uva, T. 2006, "Latent class models for utilisation of health care," *Health Economics* 15: 329-343.

Barlow, R.E. and F. Proschan 1965, *Mathematical theory of reliability*, John Wiley: New York.

Barmby, T. and J. Doornik 1989, "Modelling trip frequency as a Poisson variable," *Journal of Transport Economics and Policy* 23(3): 309-315.

Barmby, T., M. Nolan and R. Winkelmann 2001, "Contracted workdays and absence," *Manchester School*, 69(3), 269-275.

Barron, D.N. 1992, "The analysis of count data: Overdispersion and autocorrelation," in P. Marsden (ed.) *Sociological Methodology 1992*, Blackwell: Cambridge, MA, 179-220.

Bates, G. and J. Neyman 1951, "Contributions to the theory of accident proneness. II: True or false contagion," *University of California Publications in Statistics*, 215-253.

Bauer, T., A. Million, R. Rotte, and K.F. Zimmermann 1998, "Immigrant labor and workplace safety: a bivariate count data approach," mimeo.

Becker, G.S. 1962, "Investment in human capital: A theoretical analysis," *Journal of Political Economy* 70: 9-49.

Beckmann, M. 2002, "Wage compression and firm-sponsored training in Germany: Empirical evidence for the Acemoglu-Pischke model from a zero-inflated count data model," *Konjunkturpolitik* 48 (2/3).

Berglund, E. and K. Brännäs 1995, "Entry and exit of plants - a study based on Swedish panel count data," Umea Economic Studies No. 374.

Berkhout, P. and E. Plug 2004, " A bivariate poisson count data model using conditional probabilities," *Statistica Neerlandica* 58, 349-364.

Blundell, R., R. Griffith and J. van Reenen 1995, "Dynamic count data models of technological innovation," *Economic Journal* 105: 333-344.

Blundell, R., R. Griffith and J. van Reenen 1999, "Market share, market value and innovation in a panel of British manufacturing firms," *Review of Economic Studies* 66: 529-554.

Blundell, R., R. Griffith and F. Windmeijer 2002, "Individual effects and dynamics in count data models," *Journal of Econometrics* 108: 113-131.

Böckenholt, U. 1999, "Mixed INAR(1) Poisson regression models: analyzing heterogeneity and serial dependence in longitudinal count data," *Journal of Econometrics* 89: 317-338.

Böhning, D., E. Dietz, P. Schlattmann, L. Mendonca, and U. Kirchner 1999, "The zero-inflated Poisson model and the decayed, missing and filled teeth index in dental epidemiology," *Journal of the Royal Statistical Society*, Ser. A., 162: 195-209.

Börsch-Supan, A. 1990, "Education and its double-edged impact on mobility," *Economics of Education Review* 9(1): 39-53.

Boes, S. and R. Winkelmann 2006, "Ordered response models," *Advances in Statistical Analysis* 90: 165-179.

Booth, A.L., W. Arulampalam, and P. Elias 1997, "Modeling work related training and training effects using count data techniques," Centre for Economic Policy Research Discussion Paper No 1582.

Bortkiewicz, L. von 1898, *Das Gesetz der Kleinen Zahlen*, Teubner: Leipzig.

Bourlange, D. and C. Doz 1988, "Pseudo-maximum de vraisemblance: experiences de simulations dans le cadre d'un modele de Poisson," *Annales d'Economie et de Statistique* 10: 139-176.

Bowman, K.O., and Shenton, L.R. 1988, *Properties of Estimators for the Gamma Distribution*, New York: Marcel Dekker.

Brännäs, K. 1992a, "Finite sample properties of estimators and tests in Poisson regression models," *Journal of Statistical Simulation and Computation* 41: 229-241.

Brännäs, K. 1992b, "Limited dependent Poisson regression," *The Statistician* 41, 413-423.

Brännäs, K. 1994, "Estimating and testing in integer valued AR(1) models," Umea Economic Studies No. 335.

Brännäs, K. 1995a, "Explanatory variables in the INAR(1) model, " Umea Economic Studies No. 381.

Brännäs, K. 1995b, "Prediction and control for a time-series count data model," *International Journal of Forecasting* 11: 263-270.

Brännäs, K. and J. Hellström 2002, "Generalized integer-valued autoregression," *Econometric Reviews* 20: 425-43.

Brännäs, K. and P. Johansson 1994, "Time series count data regression," *Communications in Statistics: Theory and Methods* 23: 2907-2925.

Brännäs, K. and P. Johansson 1996, "Panel data regression for counts," *Statistical Papers* 37(3): 191-213.

Brännäs, K. and G. Rosenqvist 1994, "Semiparametric estimation of heterogenous count data models," *European Journal of Operational Research* 76: 247-258.

Breslow, N. 1990, "Tests of hypotheses in overdispersed Poisson regression and other quasi-likelihood models," *Journal of the American Statistical Association* 85: 565-571.

Brockett, P.L., L.L. Golden and H.H. Panjer 1996, "Flexible purchase frequency modeling," *Journal of Marketing Research*, 33: 94-107.

Broek, J. van den 1995, "A score test for zero inflation in a Poisson distribution," *Biometrics* 51: 738-743.

Buck, A.J. 1984, "Modelling a Poisson process: strike frequency in Great Britain," *Atlantic Economic Journal* 12(1):60-64.

Butler, J.S. and R. Moffitt 1982, "A computationally efficient quadrature procedure for the one-factor multinomial probit model," *Econometrica* 50: 761-764.

Cameron, A.C. and P. Johansson 1997, "Count data regression using series expansions: with applications," *Journal of Applied Econometrics* 12(3): 203-223.

Cameron, A.C. and P.K. Trivedi 1986, "Econometric models based on count data: comparisons and applications of some estimators and tests," *Journal of Applied Econometrics* 1: 29-53.

Cameron, A.C., P.K. Trivedi, F. Milne, and J. Piggott 1988, "A microeconometric model of the demand for health care and health insurance in Australia," *Review of Economic Studies* LV: 85-106.

Cameron, A.C. and P.K. Trivedi 1990, "Regression-based tests for overdispersion in the Poisson model," *Journal of Econometrics* 46: 347-364.

Cameron, A.C. and P.K. Trivedi 1998, *Regression Analysis of Count Data*, Cambridge University Press.

Cameron, A.C. and F.A.G. Windmeijer 1996, "R-squared measures for count data regression models with applications to health-care utilization," *Journal of Business and Economic Statistics* 14: 209-220.

Caudill, S.B. and F.G. Mixon 1995, "Modeling household fertility decisions: Estimation and testing of censored regression models for count data," *Empirical Economics* 20: 183-196.

Chamberlain, G. 1992, "Comment: Sequential moment restrictions in panel data," *Journal of Business and Economic Statistics* 10: 20-26.

Chappell, W.F., M.S. Kimenyi and W.J. Mayer 1990, "A Poisson probability model of entry and market structure with an application to U.S. industries during 1972-77," *Southern Economic Journal* 56: 918-927.

Chatfield, C., A.S.C. Ehrenberg, and G.J. Goodhardt 1966, "Progress on a simplified model of stationary purchasing behaviour" (with discussion), *Journal of the Royal Statistical Society* A 129: 317-367.

Chesher, A. 1984, "Testing for neglected heterogeneity," *Econometrica* 52: 865-872.

Chernoff, H. 1954, "On the distribution of the likelihood ratio," *Annals of Mathematical Statistics* 25: 573-578.

Chib, S. and E. Greenberg 1995, "Understanding the Metropolis-Hastings algorithm," *The American Statistician* 49: 327-335.

Chib, S. and E. Greenberg 1996, "Markov Chain Monte Carlo simulation methods in econometrics," *Econometric Theory* 12: 409-431.

Chib, S., E. Greenberg, and R. Winkelmann 1998, "Posterior simulation and bayes factors in panel count data models," *Journal of Econometrics* 86: 33-54.

Chib, S. and R. Winkelmann 2001, "Markov Chain Monte Carlo analysis of correlated count data," *Journal of Business and Economic Statistics* 19: 428-435.

Cincera, M. 1997, "Patents, R&D, and technological spillovers at the firm level: some evidence from econometric count models for panel data," *Journal of Applied Econometrics* 12(3): 265-80.

Congdon, P. 1989, "Modelling migration flows between areas: An analysis for London using the Census and OPCS Longitudinal Study," *Regional Studies* 23(2): 87-103.

Consul, P.C. 1989, *Generalized Poisson distributions*, Marcel Dekker: New York.

Consul, P.C. and F. Famoye 1992, "Generalized Poisson regression model," *Communications in Statistics - Theory and Methods* 21(1): 89-109.

Cox, D.R. 1961, "Tests of separate families of hypotheses," *Proc. Fourth Berkeley Symposium on Mathematical Statistics and Probability* 1: 105-123.

Cox, D.R. 1962, *Renewal Theory*, John Wiley: New York.

Cox, D.R. 1972, "Regression models and life tables," *Journal of the Royal Statistical Society* B 34: 187-202.

Cragg, J.G. 1971, "Some statistical models for limited dependent variables with application to the demand for durable goods," *Econometrica* 39: 829844.

Cramer, J.S. 1986, *Econometric Applications of Maximum Likelihood Methods*, Cambridge University Press: New York.

Creel, M.D. and J.B. Loomis 1990, "Theoretical and empirical advantages of truncated count data estimators for analysis of deer hunting in California," *American Journal of Agricultural Economics* 72(2): 434-441.

Crépon, B. and E. Duguet 1997a, "Estimating the innovation function from patent numbers: GMM on count panel data," *Journal of Applied Econometrics* 12(3): 243-263.

Crépon, B. and E. Duguet 1997b, "Research and development, competition and innovation pseudo maximum likelihood and simulated maximum likelihood methods applied to count data models with heterogeneity," *Journal of Econometrics* 79(2): 355-378.

Crouch, E.A.C. and D. Spiegelmann 1990, "The evaluation of integrals of the form $\int f(t)\exp(-t^2)dt$: application to logistic-normal models," *Journal of the American Statistical Association* 85: 464-469.

Davidson, R. and J. G. MacKinnon 1993, *Estimation and Inference in Econometrics*, Oxford University Press: Oxford.

Davutyan, N. 1989, "Bank failures as Poisson variates," *Economics Letters* 29(4): 333-338.

Dean, C. and J.F. Lawless 1989, "Tests for detecting overdispersion in Poisson regression models," *Journal of the American Statistical Association* 84: 467-472.

Dean, C., J.F. Lawless and G.E. Willmot 1989, "A mixed Poisson-inverse Gaussian regression model," *Canadian Journal of Statistics* 17(2): 171-181.

Deb, P. and P.K. Trivedi 1997, "Demand for medical care by the elderly: a finite mixture approach," *Journal of Applied Econometrics* 12(3): 313-336.

Deb, P. and P.K. Trivedi 2002, "The structure of demand for health care: latent class versus two-part models," *Journal of Health Economics* 21: 601-625.

Deb, P. and P.K. Trivedi 2006, "Specification and simulated likelihood estimation of a non-normal treatment-outcome model with selection: Application to health care utilization," *Econometrics Journal* 9: 307-331.

DeGroot, M.H. 1986, *Probability and Statistics*, Addison-Wesley: Reading, Mass.

Delgado, M.A. and T.J. Kniesner 1997, "Count data models with variance of unknown forms: an application to a hedonic model of worker absenteeism," *Review of Economics and Statistics* 79(1): 41-49.

Denuit, M. and P. Lambert 2005, "Constraints on concordance measures in bivariate discrete data," *Journal of Multivariate Analysis* 93: 4057.

Deutsches Institut für Wirtschaftsforschung (ed.) 1990, "Das Sozio-oekonomische Panel (SOEP), Benutzerhandbuch, Version 4 - Oktober/90," Berlin.

Dey, D.K. and Y. Chung 1992, "Compound Poisson distributions: properties and estimation," *Communications in Statistics – Theory and Methods* 21: 3097-3121.

Diggle, P., K.-Y. Liang, and S. L. Zeger 1995, *Analysis of Longitudinal Data*, Oxford, Oxford University Press.

Dillon, W.R. and S. Gupta 1996, "A segment-level model of category volume and brand choice," *Marketing Science* 15: 38-59.

Dionne, G. et al. 1994, "Medical conditions, risk exposure, and truck drivers' accidents," Université de Cergy-Pontoise / CNRS-Thema Discussion Paper 9409.

Dionne, G. et al. 1995, "Medical conditions and the severity of commercial motor vehicle drivers' road accidents," University of Montreal Department of Economics Discussion Paper No. 9527.

Dionne, G., R. Gagné, F. Gagnon, and C. Vanasse 1997, "Debt, moral hazard and airline safety: an empirical evidence," *Journal of Econometrics* 79(2): 379-402.

Dionne, G. and C. Vanasse 1992, "Automobile insurance ratemaking in the presence of asymmetrical information," *Journal of Applied Econometrics* 7(2): 149-165.

Doorslaer, E. van, X. Koolman, and A.M. Jones 2004, "Explaining income-related inequalities in doctor utilisation in Europe: a decomposition approach," *Health Economics* 13: 629-647.

Duan, N., W.G. Manning, C.N. Morris, and J.P. Newhouse 1983, "A comparison of alternative models for the demand for medical care," *Journal of Business and Economic Statistics* 1: 115-126.

Duijn, M.A.J. van and U. Böckenholt 1995, "Mixture models for the analysis of repeated count data, " *Applied Statistics* 44: 473-485.

Ebmer, R. 1990, "Placement service and offer arrival rates," *Economics Letters* 34: 289-294.

Efron, B. 1986, "Double exponential families and their use in generalized linear regression," *Journal of the American Statistical Association* 81: 709721.

Engle, R.F., D.F. Hendry, and J.-F. Richard 1983, "Exogeneity," *Econometrica* 51: 277-304.

Englin, J. and J.S. Shonkwiler 1995, "Estimating social welfare using count data models: an application to long-run recreation demand under conditions of endogenous stratification and truncation," *Review of Economics and Statistics* 77: 104-112.

Evans, W. N. and R. M. Schwab 1995, "Finishing high school and starting college: do catholic schools make a difference?," *Quarterly Journal of Economics* 90: 941-974.

Faddy, M.J. 1997, "Extended Poisson process modelling and analysis of count data, " *Biometric Journal* 39: 431-440.

Famoye, F. 1993, "Restricted generalized Poisson regression," *Communications in Statistics – Theory and Methods* 22: 1335-1354.

Farrow, S. 1991, "Does analysis matter? Economics and planning in the department of interior," *Review of Economics and Statistics* 73: 172-176.

Feinstein, J.S. 1989, "The safety regulations of U.S. nuclear power plants: violations, inspections, and abnormal occurrences," *Journal of Political Economy* 97: 115-154.

Feller, W. 1968, *An introduction to probability theory and its applications* Vol.1, 3rd ed., John Wiley: New York.

Feller, W. 1971, *An introduction to probability theory and its applications* Vol.2, 2nd ed., John Wiley: New York.

Firth, D. 1991, "Generalized Linear Models," in: Hinkley, D.V., N. Reid and E.J. Snell (eds.) *Statistical Theory and Modelling*, Chapman and Hall: London.

Firth, D. 1992, "Bias reduction, the Jeffreys prior, and GLIM," in: L. Fahrmeir et al. (eds.) *Advances in GLIM and statistical modelling: proceedings of the GLIM92 and the 7th International Workshop on Statistical Modelling, Munich, 13-17 June 1992.* Springer: New York, 91-100.

Flowerdew, R. and M. Aitkin 1982, "A method of fitting the gravity model based on the Poisson distribution," *Journal of Regional Science* 22(2): 191-202.

Freund, D.A., T.J. Kniesner and A.T. LoSasso 1996, "How managed care affects medicaid utilization - a synthetic difference-in-differences zero-inflated count data model," CentER Discussion Paper No 9640.

Freund, D.A., T.J. Kniesner and A.T. LoSasso 1999, "Dealing with the common econometric problems of count data with excess zeros, endogenous treatment effects, and attrition bias," *Economics Letters* 62: 7-12.

Frome, E.L , M.H. Kutner and J.J. Beauchamp 1973, "Regression analysis of Poisson-distributed data," *Journal of the American Statistical Association* 68: 935-940.

Gallant, A.R. and D. W. Nychka 1987, "Semi-nonparametric maximum likelihood estimation," *Econometrica* 55: 363-390.

Gameren, E. van and I. Woittiez 2002, "Determinants of the demand for home care: the effect of supply restrictions," mimeo. Central Planing Bureau.

Gamerman, D. 1997, *Markov Chain Monte Carlo*, Chapman & Hall: London.

Ganio, L.M. and D.W. Schafer 1992, "Diagnostics for overdispersion," *Journal of the American Statistical Association* 87: 795-804.

Geil, P., A. Million, R. Rotte and K.F. Zimmermann 1997, "Economic incentives and hospitalization in Germany," *Journal of Applied Econometrics* 12(3): 295-311.

Gerdtham, U.G. 1997, "Equity in health care utilization: further tests based on hurdle models and Swedish micro data," *Health Economics* 6(3): 303-319.

Gilbert, C.L. 1982, "Econometric models for discrete (integer valued) economic processes," in: E.G. Charatsis (ed.) *Selected Papers on Contemporary Econometric Problems*, The Athens School of Economics and Business science.

Goldberger, A.S. 1968, "The interpretation and estimation of Cobb-Douglas functions," *Econometrica* 35: 464-472.

Good, D.H., and M.A. Pirog-Good 1989, "Models for bivariate count data with an application to teenage delinquency and paterny," *Sociological Methods & Research* 17: 409-431.

Gourieroux, C. 1989, *Econométrie des Variables Qualitatives*, 2nd ed., Economica: Paris.

Gourieroux, C., A. Monfort and A. Trognon 1984a, "Pseudo maximum likelihood methods: Theory," *Econometrica* 52: 681-700.

Gourieroux, C., A. Monfort and A. Trognon 1984b, "Pseudo maximum likelihood methods: applications to Poisson models," *Econometrica* 52: 701-721.

Gourieroux, C. and A. Monfort 1989, *Statistique et Modèles Econométriques*, Vol. 1 and 2, Economica: Paris.

Gourieroux, C. and A. Monfort 1993, "Simulation-based inference: a survey with special reference to panel data models," *Journal of Econometrics*, 59: 5-33.

Gourieroux, C. and M. Visser 1997, "A count data model with unobserved heterogeneity," *Journal of Econometrics* 79(2): 247-268.

Greene, W.H. 1995, *LIMDEP 7.0 User's Manual*, Econometric Software Inc.: Bellport, NY.

Greene, W.H. 1998, "Sample selection in credit-scoring models," *Japan and the World Economy* 10(3): 299-316.

Greene, W.H. 2000, *Econometric analysis*, 4th ed., Prentice Hall: New York.

Greene, W.H. 2007, "Functional forms for the negative binomial model for count data," *Economics Letters*, doi: 10.1016/j.econlet.2007.10.015

Greenwood, M. and G.U. Yule 1920, "An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the ocurrence of multiple attacks of disease or of repeated accidents," *Journal of the Royal Statistical Society* A 83: 255-279.

Grogger, J.T. 1990a, "The deterrent effect of capital punishment: an analysis of daily homicide counts," *Journal of the American Statistical Association* 85: 295-303.

Grogger, J.T. 1990b, "A simple test for exogeneity in probit, logit, and Poisson regression models," *Economics Letters* 33: 329-332.

Grogger, J.T. and R.T. Carson 1991, "Models for truncated counts," *Journal of Applied Econometrics* 6: 225-238.

Grootendorst, P.V. 1995, "A comparison of alternative models of prescription drug utilization," *Health Economics* 4: 183-198.

Guldberg, A. 1931, "On discontinuous frequency-functions and statistical series," *Skandinauisk Aktuarietionskrift* 14: 167-197.

Guo, J.Q. and T. Li 2001, "Simulation-based estimation of the structural errors-in-variables negative binomial regression model with an application," *Annals of Economics and Finance* 2: 101-122.

Guo, J.Q. and T. Li 2002, "Poisson regression models with errors-in-variables: implication and treatment," *Journal of Statistical Planning and Inference*, 104: 391-401.

Guo, J. Q. and P.K. Trivedi 2002, Flexible parametric models for long-tailed patent count distributions, *Oxford Bulletin of Economics and Statistics*, 63: 63-82.

Gupta, S. 1988, "Impact of sales promotions on when, what, and how much to buy," *Journal of Marketing Research* 25: 342-355.

Gurmu, S. 1991, "Tests for detecting overdispersion in the positive Poisson regression model," *Journal of the Business and Economics Statistics* 9: 215-222.

Gurmu, S. 1997, "Semi parametric estimation of hurdle regression models with an application to medicaid utilization," *Journal of Applied Econometrics* 12(3): 225-243.

Gurmu, S. 1998, "Generalized hurdle count data regression models," *Economics Letters* 58(3): 263-268.

Gurmu, S. and J. Elder 1998, "Estimation of multivariate count regression models with applications to health care utilization," mimeo.

Gurmu, S. and J. Elder 2000, "Generalized bivariate count data regression models," *Economics Letters* 68: 31-36

Gurmu, S., P. Rilstone and S. Stern 1998, "Semiparametric estimation of count regression models," *Journal of Econometrics* 88(1): 123-150.

Gurmu, S. and P.K. Trivedi 1992, "Overdispersion tests for truncated Poisson regression models," *Journal of Econometrics* 54: 347-370.

Gurmu, S. and P.K. Trivedi 1996, "Excess zeros in count models for recreational trips," *Journal of Business and Economic Statistics* 14(4): 469-477.

Haab, T.C. and K.E. McConnell 1996, "Count data models and the problem of zeros in recreation demand analysis," *American Journal of Agricultural Economics* 78: 89-102.

Hahn, R.W. and J.E. Prieger 2006, "The impact of driver cell phone use on accidents," *Advances in Economic Analysis & Policy* 6(1), Article 9.

Hall, R.E. 1982, "The importance of lifetime jobs in the U.S. Economy," *American Economic Review* 72: 716-724.

Hansen, L.P. 1982, "Large sample properties of generalized method of moment estimators," *Econometrica* 50: 1029-1054.

Hastie, T. and R. Tibshirani 1986, "Generalized additive models", *Statistical Science* 1: 297-318.

Hausman, J.A. 1978, "Specification tests in econometrics," *Econometrica* 46: 1251-1271.

Hausman, J.A. , B.H. Hall and Z. Griliches 1984, "Econometric models for count data with an application to the Patents-R&D relationship," *Econometrica* 52: 909-938.

Heckman, J.J. 1978, "Dummy endogenous variables in simultaneous equation systems," *Econometrics* 46: 931-959.

Heckman, J.J. 1979, "Sample selection bias as a specification error," *Econometrica* 47: 153-161.

Heckman, J.J. 1981, "Heterogeneity and state dependence," in: Rosen, S. (ed.) *Studies in Labor Markets* University of Chicago Press: Chicago.

Heckman, J.J. and G. Borjas 1980, "Does unemployment cause future unemployment? Definitions, questions and answers from a continuous time model for heterogeneity and state dependence," *Economica* 47: 247-283.

Heckman, J.J. and B. Singer 1984, "A method for minimizing the impact of distributional assumptions in econometric models for duration data," *Econometrica* 52: 271-320.

Hellström, J. 2002, "Count data modelling and tourism demand," Umea Economic Studies No. 584.

Hinde, J. 1982, "Compound Poisson regression models," in: R. Gilchrist (ed.) *GLIM 1982: Proceedings of the international conference on Generalized linear models*, Springer: Berlin.

Hinde, J. 1992, "Choosing between non-nested models: A simulation approach," in: L. Fahrmeir et al. (eds.) *Advances in GLIM and statistical modelling: proceedings of the GLIM92 and the 7th International Workshop on Statistical Modelling, Munich, 13-17 June 1992.* Springer: New York, 119-124.

Hinkley, D.V. and N. Reid 1991, "Statistical Theory," in: Hinkley, D.V., N. Reid and E.J. Snell (eds.) *Statistical Theory and Modelling*, Chapman and Hall: London.

Hogg, R.V. and A.T. Craig 1978, *Introduction to Mathematical Statistics*, 4th ed., Macmillan: New York.

Hsiao, C. 1986, *Analysis of Panel Data*, New York: Cambridge University Press.

Jaggia, S. and S. Thosar 1993, "Multiple bids as a consequence of target management resistance: a count data approach," *Review of Quantitative Finance and Accounting* 3: 447-457.

Jensen, E.J. 1987, "Research expenditures and the discovery of new drugs," *Journal of Industrial Economics* 36(1): 83-95.

Jimenéz-Martin, S. , J.M. Labeaga and M. Martinez-Granado 2002, "Latent class versus two-part models in the demand for physician services across the European Union," *Health Economics* 11: 301-321.

Johansson, P. and K. Brännäs 1995, "Work absence within the household," in: C.S. Forbes, P. Kofman, and T.R.L. Fry (eds.) *Proceedings of the 1995 Econometrics Conference at Monash*, Department of Economics: Melbourne.

Johansson, P. and M. Palme 1996, "Do economic incentives affect work absence? empirical evidence using Swedish micro data," *Journal of Public Economics* 59(2): 195-218.

Johnson, N.L. and S. Kotz 1969, *Distributions in statistics: discrete distributions*, Wiley: New York.

Johnson, N.L. and S. Kotz 1970, *Distributions in statistics: continuous distributions*, Houghton-Mifflin: Boston.

Johnson, N.L., S. Kotz and N. Balakrishnan 1994, *Continuous Univariate Distributions*, Vol. 1, 2nd ed., Wiley: New York.

Jørgensen, B., Lundbye-Christensen, S., Xue-Kun Song, P. and Sun, L. 1999, "A state space models for multivariate longitudinal count data," *Biometrika* 86: 169-181.

Jorgenson, D.W. 1961, "Multiple regression analysis of a Poisson process," *Journal of the American Statistical Association* 56: 235-245.

Jovanovic, B. 1979a, "Job matching and the theory of turnover," *Journal of Political Economy* 87: 972-990.

Jovanovic, B. 1979b, "Firm specific capital and turnover," *Journal of Political Economy* 87: 1246-1260.

Jung, R. 1999, *Zeitreihenanalyse für Zähldaten: Eine Untersuchung ganzzahliger Autoregressiver-Moving-Average-Prozesse*, (Reihe: Quantitative Ökonomie, Vol. 100) Eul: Lohmar ; Köln.

Jung, R. and R. Liesenfeld 2001, "Estimating time series models for count data using efficient importance sampling," *Allgemeines Statistisches Archiv* 85: 387-407.

Jung, R. and R. Winkelmann 1993, "Two aspects of labor mobility: A bivariate Poisson regression approach," *Empirical Economics* 18: 543-556.

Kahn, M.E. 2005, "The death toll from natural disasters: the role of income, geography, and institutions," *Review of Economics and Statistics* 87: 271-284.

Kalwij, A. 2000, "The effects of female employment status on the presence and number of children," *Journal of Population Economics* 13: 221-240.

Karlis, D. and E. Xekalaki 2005, "Mixed Poisson distributions," *International Statistical Review* 73: 3558.

Kelly, M. 2000, "Inequality and crime," *Review of Economics and Statistics* 82: 530-539.

Kenkel, D.S. and J.V. Terza 2001, "The effect of physician advice on alcohol consumption: count regression with an endogenous treatment effect," *Journal of Applied Econometrics* 16: 165-184.

Kennan, J. 1985, "The duration of contract strikes in U.S. manufacturing," *Journal of Econometrics* 28: 5-28.

Kennedy, P. 1981, "Estimation with correctly interpreted dummy variables in semilogarithmic equations," *American Economic Review* 71: 801.

Kennedy, W.J. and J.E. Gentle 1980, *Statistical Computation*, Marcel Dekker: New York.

King, G. 1988, "Statistical models for political science event counts: bias in conventional procedures and evidence for the exponential Poisson regression model," *American Journal of Political Science* 32: 838-862.

King, G. 1989a, "A seemingly unrelated Poisson regression model," *Sociological Methods & Research* 17: 235-255.

King, G. 1989b, "Variance specification in event count models: from restrictive assumptions to a generalized estimator," *American Journal of Political Science* 33: 762-784.

King, G. 1989c, *Unifying political methodology: the likelihood theory of statistical inference*, Cambridge University Press: Cambridge.

Knuth, D.E. 1969, *The Art of Computer Programing;* Vol.2 *Seminumerical Analysis*. Addison-Wesley: Reading, Mass.

Kocherlakota, S. and K. Kocherlakota 1992, *Bivariate discrete distributions*, Marcel Dekker: New York.

Koenker, R. and K. F. Hallok 2001, "Quantile regression," *Journal of Economic Perspectives* 15: 143-156.

Kostiuk, P.F. and D.A. Follmann 1989, "Learning curves, personal characteristics, and job performance," *Journal of Labor Economics* 7(2): 129-146.

Kozumi, H. 2002, "A Bayesian analysis of endogenous switching models for count data," *Journal of the Japan Statistical Society* 32: 141154.

Kulaserka, K.B. and D.W. Tonkyn 1992, "A new discrete distribution with applications to survival, dispersal and dispersion," *Communications in Statistics - Simulations and Computations* 21(2): 499-518.

Laird, N. 1978, "Nonparametric maximum likelihood estimation of a mixing distribution," *Journal of the American Statistical Association* 73: 805-811.

Lakshminarayana, J., S.N.N. Pandit and K.S. Rao 1999, "On a bivariate Poisson distribution," *Communications in Statistics – Theory and Methods* 28: 267-276.

Lambert, D. 1992,"Zero-inflated Poisson regression with an application to defects in manufacturing," *Technometrics* 34: 1-14.

Lancaster, T. 1990, *The econometric analysis of transition data.* Cambridge University Press: Cambridge, U.K.

Lawless, J.F. 1987a, "Regression methods for Poisson process data," *Journal of the American Statistical Association* 82: 808-815.

Lawless, J.F. 1987b, "Negative binomial and mixed Poisson regression," *The Canadian Journal of Statistics* 15(3): 209-225.

Lazear, E. 1990, "The job as a concept," The Hoover Institution Woking Paper E-90-24.

Lee, L.-F. 1983, "Generalized econometric models with selectivity, " *Econometrica* 51: 507-512.

Lee, L.-F. 1986, "Specification test for Poisson regression models," *International Economic Review* 27: 687-706.

Lee, L.-F. 1996, "Specification and estimation of count data regression and sample selection models – a counting process and waiting time approach," hong kong university of sciece and technology, Department of Economics Working Paper No. 96/19.

Lindsay, B.G. 1983, "The geometry of mixture likelihoods," *Annals of Statistics* 11: 86-94.

List, J.A. 2002, "Determinants of securing academic interviews after tenure denial: Evidence from a zero inflated Poisson model," *Applied Economics* 33: 1423-1431.

Long and Bowyer 1953, "The influence of earnings on the mobility of labour," *Yorkshire Bulletin of Economic and Social Research* 5: 81-87.

Maasoumi, E. 1992, "Rules of thumb and pseudo science," *Journal of Econometrics* 53: 1-4.

Machado, J.A.F. and J.M.C. Santos Silva 2005, "Quantiles for Counts," *Journal of the American Statistical Association* 100: 1226-1237.

Maddala, G.S. 1983, *Limited-Dependent and Qualitative Variables in Econometrics*, Cambridge University Press: Cambridge.

Manski, C.F. and S.R. Lerman 1977, "The estimation of choice probabilities from choice based samples," *Econometrica* 45: 1977-1988.

Marshall, A.W. and I. Olkin 1990, "Multivariate distributions generated from mixtures of convolution and product families, " in: H.W. Block, A.R. Sampson, and T.H. Savits (eds.) *Topis in Statistical Dependence*, 371-393, IMS Lecture Notes Monograph Series, Volume 16.

Maul, A., A.H. El-Shaarawi and J.F. Ferard 1991, "Application of negative binomial regression models to the analysis of quantal bioassay data," *Environmetrics* 2: 253-261.

Mayer, J. and R. Riphahn 2000, "Fertility assimilation of immigrants: a varying coefficient count data model," *Journal of Population Economics* 13: 241-262.

McCullagh, P. 1991, "Quasi-likelihood and estimating functions," in: Hinkley, D.V., N. Reid and E.J. Snell (eds.) *Statistical Theory and Modelling*, Chapman and Hall: London.

McCullagh, P. and J.A. Nelder 1989, *Generalized linear models*, 2nd ed., Chapman and Hall: London.

McIntosh, J. 1999, "An analysis of reproductive behavior in canada: results from an intertemporal optimizing model," *Journal of Population Economics* 12: 451-461.

McKelvey, R.D. and W. Zavoina 1975, "A statistical model for the analysis of ordinal level dependent variables," *Journal of Mathematical Sociology* 4: 103-120.

McKenzie E. 1986, "Autoregressive moving average provesses with negative binomial and geometric marginal distributions," *Advances in Applied Probability* 18: 679-705.

McKenzie E. 1988, "Some ARMA models for dependent sequences for Poisson counts," *Advances in Applied Probability* 20: 822-835.

Melkersson, M. and C. Olsson 1999, "Is visiting the dentist a good habit? analyzing count data with excess zeros and excess ones," *Umea Economic Studies* No. 492.

Melkersson, M. and D. Roth 2000, "Modeling of household fertility using inflated count data models," *Journal of Population Economics* 13: 189-204.

Merkle, L. and K.F. Zimmermann 1992, "The demographics of labor turnover: A comparison of ordinal probit and censored count data models," *Recherches Economiques de Louvain* 58: 283-306.

Michener, R. and C. Tighe 1992, "A Poisson regression model of highway fatalities," *American Economic Review* 82: 452-456.

Million, A. 1998, *Zähldatenmodelle für korrelierte abhängige Daten*, unpublished Ph.D. thesis, University of Munich.

Mincer, J. 1962, "On-the-job training: Costs, returns, and some implications," *Journal of Political Economy* 70: 50-79.

Mincer, J. and B. Jovanovic 1981, "Labor mobility and wages," in: S. Rosen (ed.) *Studies in Labor Markets*, The University of Chicago Press: Chicago, 21-62.

Miranda A. 2008, "Planed fertility and family background: a quantile regression for counts analysis," *Journal of Population Economics* 21(1).

Moffatt, P.G. 1995, "Grouped poisson regression models: theory and an application to public house visit frequency," *Communications in Statistics: Theory and Methods* 24: 2779-2796.

Moffatt, P.G. 1997a, "Exploiting a matrix identity in the computation of the efficient score test for overdispersion in the Poisson regression model," *Statistics and Probability Letters* 32: 75-79.

Moffatt, P.G. 1997b, "Global log-concavity of the likelihood in models with grouped discrete data," *Australian Journal of Statistics* 39: 105-112.

Moffatt, P.G. and S.A. Peters 2000, "Grouped zero-inflated count data models of coital frequency," *Journal of Population Economics* 13: 205-220.

Møller Danø, Anne 1999, "Unemployment and health conditions – a count data approach," mimeo.

Montalvo, J.G. 1997, "GMM estimation of count panel data models with fixed effects and predetermined instruments," *Journal of Business and Economic Statistics* 15(1): 82-89.

Mortensen, D.T. 1986, "Job search and labor market analysis," in: O. Ashenfelter and R. Layard (eds.) *Handbook of Labor Economics* Vol.II, Amsterdam, North-Holland.

Mroz, T. 1999, "A discrete factor approximations for use in simultaneous equation models: estimating the impact of a dummy endogenous variable on a continuous outcome," *Journal of Econometrics* 92: 233-274.

Mukhopadhyay, K. and P.K. Trivedi 1995, "Regression models for under-recorded count data," paper presented at the 7th World Congress of the Econometric Society.

Mullahy, J. 1986, "Specification and testing in some modified count data models," *Journal of Econometrics* 33: 341-365.

Mullahy, J. 1997a, "Instrumental variable estimation of count data models: applications to models of cigarette smoking behavior," *Review of Economics and Statistics* 79(4): 586-593.

Mullahy, J. 1997b, "Heterogeneity, excess zeros, and the structure of count data models," *Journal of Applied Econometrics* 12(3): 337-50.

Mullahy, J. 1999, "Interaction effects and difference-in-difference estimation in loglinear models," NBER Technical Working Paper No. 245.

Mullahy, J. and P.R. Portney 1990, "Air pollution, cigarette smoking and the production of respiratory health," *Journal of Health Economics* 9: 193-205.

Munkin, M.K. and P.K. Trivedi 1999, "Simulated maximum likelihood estimation of multivariate mixed-Poisson regression models, with application," *Econometrics Journal* 2: 29-48.

Murphy, K. and R. Topel 1985, "Estimation and inference in two step econometric models," *Journal of Business and Economic Statistics* 3: 370-379.

Nakamura, T. 1990, "Corrected score function for errors-in-variables models: methodology and application to generalized linear models," *Biometrika* 77: 127-137.

Nelder, J.A. and R.W. Wedderburn 1972, "Generalized linear models," *Journal of the Royal Statistical Society* A 135: 370-384.

Nguyen-Dinh, H. 1997, "A socioeconomic analysis of the determinants of fertility: the case of Vietnam," *Journal of Population Economics* 10: 251-272.

Okoruwa, A.A., J.V. Terza and H.O. Nourse 1988, "Estimating patronization shares for urban retail centers: An extension of the Poisson gravity model," *Journal of Urban Economics* 24(3): 241-259.

Ophem, H. van 1999, "A general method to estimate correlated discrete random variables," *Econometric Theory* 15(2): 228-237.

Ophem, H. van 2000, "Modeling selectivity in count data models," *Journal of Economic and Business Statistics* 18: 503-511.

Ozuna, T. and I.A. Gomez 1994, "Estimating a system of recreation demand functions using a seemingly unrelated poisson regression approach," *Review of Economics and Statistics* 76(2): 356-360.

Ozuna, T. and I.A. Gomez 1995, "Specification and testing of count data recreation demand functions," *Empirical Economics* 20(3): 543-550.

Pesaran, M.H. 1974, "On the general problem of model selection," *Review of Economic Studies* 41: 153-171.

Pfeiffer, P.E. 1978, *Concepts of probability theory*, 2nd ed., Dover: New York.

Plassmann, F. and T.N. Tideman 2001, "Does the right to carry concealed handguns deter countable crimes? only a count analysis can say," *Journal of Law and Economics* 44: 771-798.

Pohlmeier, W. and V. Ulrich 1995, "An econometric model of the two-part decision making process in the demand for health care," *Journal of Human Resources* 30: 339-361.

Praag, B.M.S. van and E.M. Vermeulen 1993, "A count-amount model with endogenous recording of observations," *Journal of Applied Econometrics* 8: 383-395.

Prieger, J.E. 2002, "Regulation, innovation and the introduction of new telecommunications services," *Review of Economics and Statistics* 84: 704-715.

Prieger, R.W. and J.E. Hahn 2006, "Are drivers who use cell phones inherently less safe?," *Applied Economics Quarterly* 53(4).

Ramaswamy, V., E.W. Anderson and W.S. DeSarbo 1996, "A disaggregate negative binomial regression procedure for count data analysis," *Management Science* 40: 405-417.

Rickard, J.M. 1988, "Factors influencing long distance rail passenger trip rates in Great Britain," *Journal of Transport Economics and Policy* 22(2): 209-233.

Riphahn, R.T., A. Wambach and A. Million 2003, "Incentive effects in the demand for health care: a bivariate panel count data estimation," *Journal of Applied Econometrics* 18: 387-405.

Rivers, D. and Q.H. Vuong 1988, "Limited information estimators and exogeneity tests for simultaneous probit models," *Journal of Econometrics* 39: 347-366.

Romeu, A. and M. Vera-Hernandez 2005, "Counts with an endogenous binary regressor: A series expansion approach," *Econometrics Journal* 8, 1-22.

Ronning, G. and R. Jung 1992, "Estimation of a first order autoregressive process with Poisson marginals for count data," in: L. Fahrmeir et al. (eds.) *Advances in GLIM and statistical modelling: proceedings of the GLIM92 and the 7th International Workshop on Statistical Modelling, Munich, 13-17 June 1992.* Springer: New York, 188-194.

Rose, N.L. 1990, "Profitability and product quality: Economic determinants of airline safety performance," *Journal of Political Economy* 98: 944-964.

Rosen, S. 1981, "Introduction," in: Rosen, S. (ed.) *Studies in Labor Markets*, University of Chicago Press: Chicago, 1-19.

Rosen, S. 1992, "Distinguished Fellow: Mincerian labor economics," *Journal of Economic Perspectives* 6(2): 157-170.

Ross, S.M. 1985, *Introduction to Probability Models*, Academic Press: New York.

Roy, A.D. 1951, "Some thoughts on the distribution of earnings," *Oxford Economic Papers* 3: 135-145.

Ruser, J. 1991, "Workers' compensation and occupational injuries and illnesses," *journal of labor economics* 9(4): 325-350.

Ruser, J. 1993, "Workers' compensation and the distribution of occupational injuries," *Journal of Human Resources* 28(3): 593-617.

Saha, A. and D. Dong 1997, "Estimating nested count data models," *Oxford Bulletin of Economics and Statistics* 59(3): 423-430.

Sander, W. 1992, "The effects of women's schooling on fertility," *Economics Letters* 40: 229-233.

Santos Silva, J.M.C. 1997a, "Unobservables in count data models for on site samples," *Economics Letters* 54(3): 217-220.

Santos Silva, J.M.C. 1997b, "Generalized Poisson regression for positive count data," *Communications in Statistics – Simulations and Computations* 26 (3).

Santos Silva, J.M.C. and F. Covas 2000, "A modified hurdle model for completed fertility," *Journal of Population Economics* 13: 173-188.

Santos Silva, J.M.C. 2001, "A score test for non-nested hypotheses with applications to discrete data models," *Journal of Applied Econometrics* 16: 577-597.

Santos Silva, J.M.C. 2003, "A note on the estimation of mixture models under endogenous sampling," *Econometrics Journal* 6: 46-52.

Santos Silva, J.M.C. and F. Windmeijer 2001, "Two-part multiple spell models for health care demand," *Journal of Econometrics* 104: 67-89.

Santos Silva, J.M.C. and S. Tenreyro 2006, "The log of gravity," *Review of Economics and Statistics* 88, 641-658.

Sapra, S. 2005, "A regression error specification test (RESET) for generalized linear models", *Economics Bulletin* 3: 1-6.

Schellhorn, M. 2001, "The effect of variable health insurance deductibles on the demand for physician visits," *Health Economics* 10: 441-456.

Schultz, T.P. 1990, "Testing the neoclassical model of family labor supply and fertility," *Journal of Human Resources* 25(4): 599-634.

Schwalbach, J. and K.F. Zimmermann 1991, "A Poisson model of patenting and firm structure in Germany," in: Z. Acs and D. Audretsch (Eds.) *Innovation and Technological Change*, Harvester Wheatsheaf, New York et al., 109-120.

Shaked, M. 1980, "On mixtures from exponential families," *Journal of the Royal Statistical Society Series B* 42: 192-198.

Shaw, D. 1988, "On-site samples regression," *Journal of Econometrics* 37: 211-223.

Shonkwiler, J.S. and T.R. Harris 1996, "Rural retail business thresholds and interdependencies," *Journal of Regional Science* 36: 617-630.

Shonkwiler, J.S. and W.S. Shaw 1996, "Hurdle count data models in recreation demand analysis," *Journal of Agricultural and Resource Economics* 21(2): 210-219.

Silcock 1954, "The phenomenon of labour turnover," *Journal of the Royal Statistical Society* B 16: 429-440.

Simar, L. 1976, "Maximum likelihood estimation of a compound Poisson process," *Annals of Statistics* 4: 1200-1209.

Skvoretz, J. 1984, "Career mobility as a Poisson process," *Social Science Research* 13: 198-220.

Smith, T.E. 1987, "Poisson gravity models of spatial flows," *Journal of Regional Science* 27(3): 315-340.

Statistisches Bundesamt (ed.) 1985, "Statistisches Jahrbuch 1985 für die Bundesrepublik Deutschland," Wiesbaden.

StataCorp. (1997), "Stata 5.0 Reference Manual," StataPress: College Station.

StataCorp. (1999), *Stata Statistical Software: Release 6.0.* Stata Corporation: College Station, TX.

Terza, J.V. 1985, "A Tobit type estimator for the censored Poisson regression model," *Economics Letters* 18: 361-365.

Terza, J.V. 1998, "Estimating count data models with endogenous switching: sample selection and endogenous treatment effects," *Journal of Econometrics* 84(1): 129-154.

Tomlin, K.M. 2000, "The effects of model specification on foreign direct investment models: an application of count data models," *Southern Economic Journal* 67: 460-468.

Topel, R.H. and M.P. Ward 1992, "Job mobility and the careers of young men," *Quarterly Journal of Economics* 106(2): 439-479.

Trivedi, P.K. and D.M. Zimmer 2007, "Copula modeling: an introduction for practitioners," *Foundations and Trends in Econometrics* 1.

Veall, M.R. and K.F. Zimmermann 1992, "Performance measures from prediction-realization tables," *Economics Letters* 39: 129-134.

Vera-Hernandez, M. 1999, "Duplicate coverage and demand for health care: the case of Catalonia," *Health Economics* 8: 579-598.

Vistnes, J.P. 1997, "Gender differences in days lost from work due to illness," *Industrial and Labor Relations Review* 50(2): 304-323.

Vuong, Q.H. 1989, "Likelihood ratio tests for model selection and non-nested hypotheses," *Econometrica* 57(2): 307-333.

Wagner, G.G., R.V. Burkhauser and F. Behringer 1993, "The English language public use file of the German Socio-Economic Panel," *Journal of Human Resources* 28: 429-433.

Wang, P., M.L. Puterman, I.M. Cockburn, and N. Lee 1997, "Mixed Poisson regression model with covariate dependent rates," *Biometrics* 52: 381-400.

Wang, P., I.M. Cockburn, and M.L. Puterman 1998, "Analysis of patent data: a mixed Poisson regression model approach," *Journal of Business and Economic Statistics* 16(1): 27-41.

Wang, W. and F. Famoye 1997, "Modeling household fertility decisions with generalized Poisson regression," *Journal of Population Economics* 10(3): 273-283.

Wedel, M., W.S. Desarbo, J.R. Bult and V. Ramaswamy 1993, "A latent class Poisson regression model for heterogeneous count data," *Journal of Applied Econometrics* 8: 397-411.

Weiss, A. 1999, "A simultaneous binary choice / Poisson regression model with an application to credit card approvals," mimeo.

White, H. 1982, "Maximum likelihood estimation of misspecified models," *Econometrica* 50(1): 1-25.

Williams, D.A. 1970, "Discriminating between regression models to determine the pattern of enzyme synthesis in sychronous cell cultures," *Biometrics* 28: 23-32.

Wilson, P.W. 1992, "Count data models without mean-variance restrictions," presented at the European Meeting of the Econometric Society, Brussels

Windmeijer, F.A.G. 2000, "Moment conditions for fixed effects count data models with endogenous regressors," *Economics Letters* 68: 21-24.

Windmeijer, F.A.G. 2008, "GMM for panel count data models," in: M. Laszlo and P. Sevestre (eds.) *The Econometrics of Panel Data*, Springer.

Windmeijer, F.A.G. and J.M.C. Santos Silva 1997, "Endogeneity in count data models: an application to demand for health care," *Journal of Applied Econometrics* 12(3): 281-294.

Winkelmann, R. 1995, "Duration dependence and dispersion in count data models," *Journal of Business and Economic Statistics* 13: 467-474.

Winkelmann, R. 1996a, "A count data model for gamma waiting times," *Statistical Papers* 37: 177-187.

Winkelmann, R. 1996b, "Markov chain Monte Carlo analysis of underreported count data with an application to worker absenteeism," *Empirical Economics* 21: 575-587.

Winkelmann, R. 1997, "Poisson regression with endogenous reporting," in: P. Bardsley and V.L. Martin (eds.), Proceedings of the Econometric Society Australasian Meeting 2nd-4th July 1997 volume 4: Microeconometrics, 385-398.

Winkelmann, R. 1998, "Count data models with selectivity," *Econometric Reviews* 17: 339-359.

Winkelmann, R. 1999, "Wages, firm size, and absenteeism," *Applied Economics Letters* 6: 337-341.

Winkelmann, R. 2000, "Seemingly unrelated negative binomial regression," *Oxford Bulletin of Economics and Statistics*, 62(4), 553-560.

Winkelmann, R. 2001, "Correctly interpreting the results from a log-linear regression under heteroskedasticty - methods and an application to the rel-

ative wages of immigrants," *Jahrbücher für Nationalökonomie und Statistik* 221: 418-431.

Winkelmann, R. 2004a, "Co-payments for prescription drugs and the demand for doctor visits - evidence from a natural experiment," *Health Economics* 13: 1081-1089.

Winkelmann, R. 2004b, "Health care reform and the number of doctor visits - an econometric analysis," *Journal of Applied Econometrics* 19: 455-472.

Winkelmann, R. 2006, "Reforming health care: evidence from quantile regressions for counts," *Journal of Health Economics* 25: 131-145.

Winkelmann, R. and K.F. Zimmermann 1991, "A new approach for modeling economic count data," *Economics Letters* 37: 139-143.

Winkelmann, R. and K.F. Zimmermann 1992a, "Robust Poisson regression," in: L. Fahrmeir et al. (eds.) *Advances in GLIM and statistical modelling: proceedings of the GLIM92 and the 7th International Workshop on Statistical Modelling, Munich, 13-17 June 1992.* Springer: New York, 201-206.

Winkelmann, R. and K.F. Zimmermann 1992b, "Recursive probability estimators for count data," in: Haag, G., U. Mueller and K.G. Troitzsch (eds.) *Economic Evolution and Demographic Change.* Springer: New York, 321-329.

Winkelmann, R. and K.F. Zimmermann 1993a, "Job separations in an efficient turnover model," in: Bunzel, H., P. Jensen and N. Westergard-Nielsen, *Panel Data and Labour Market Dynamics* North-Holland: Amsterdam, 107-122.

Winkelmann, R. and K.F. Zimmermann 1993b, "Ageing, migration and labour mobility," in: P. Johnson and K.F. Zimmermann (eds.) *Labour Markets in an Ageing Europe*, Cambridge University Press: Cambridge U.K., 255-283.

Winkelmann, R. and K.F. Zimmermann 1993c, "Poisson-Logistic Regression," Department of Economics, University of Munich, Working Paper No. 93-18.

Winkelmann, R. and K.F. Zimmermann 1994, "Count data models for demographic data," *Mathematical Population Studies* 4: 205-221.

Winkelmann, R. and K.F. Zimmermann 1995, "Recent developments in count data modeling: Theory and applications," *Journal of Economic Surveys* 9: 1-24.

Winkelmann, R. and K.F. Zimmermann 1998, "Is job stability declining in Germany? Evidence from count data models," *Applied Economics* 30: 1413-1420.

Winkelmann, R. and K.F. Zimmermann 2000, "Editorial: Fertility studies using count data models," *Journal of Population Economics* 13: 171-172.

Winkelmann, R., C. Signorino, and G. King 1995, "A correction for an underdispersed event count probability distribution," *Political Analysis* 5: 215-228.

Wooldridge, J.M. 1992, "Some alternatives to the Box-Cox regression model, *International Economic Review* 33: 935-955.

Wooldridge, J.M. 1997a, "Multiplicative panel data models without the strict exogeneity assumption," *Econometric Theory* 13: 667-679.

Wooldridge, J.M. 1997b, "Quasi-likelihood methods for count data," in: M.H. Pesaran amd P. Schmidt (eds.) *Handbook of Applied Econometrics* Vol. II, Microeconometrics, Blackwell.

Wooldridge, J.M. 1999, "Distribution-free estimation of some nonlinear panel data models," *Journal of Econometrics* 90: 77-97.

Wooldridge, J.M. 2002, "Econometric analysis of cross section and panel data," MIT Press.

Wun, L.-M. 1991, "Regression analysis for autocorrelated Poisson distributed data," *Communications in Statistics - Theory and Methods* 20(10): 3083-3091.

Yen, S.T. 1999, "Gaussian versus count-data hurdle models: cigarette consumption by women in the US," *Applied Economics Letters* 6: 73-76.

Yousry, M.A. and R.C. Srivastava 1987, "The hyper negative binomial distribution," *Biometric Journal* 29: 875-884.

Zeger, S.L. 1988, "A regression model for time series of counts," *Biometrika* 75(4): 621-629.

Zellner, A. 1962, "An efficient method of estimating seemingly unrelated regressions and tests of aggregation bias," *Journal of the American Statistical Association* 57: 348-368.

Zellner, A. 1971, *An Introduction to Bayesian Inference in Econometrics*, John Wiley: New York.

Zimmer, D. M. and P.K. Trivedi 2006, "Using trivariate copulas to model sample selection and treatment effects: application to family health care demand," *Journal of Business and Economic Statistics* 24: 63-76.

Zimmermann, K.F. and J. Schwalbach 1991, "Determinanten der Patentaktivität," *Ifo-Studien* 37: 201-227.

Author's Index

Subject Index