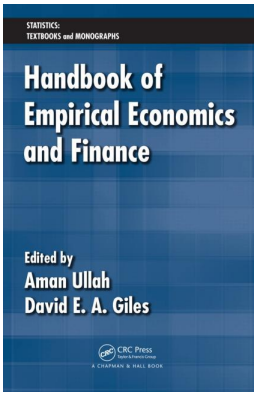

Handbook of Empirical Economics and Finance

Ullah Aman, E. A. Giles David

Recent Developments in Cross Section and Panel Count Models

Publication details
https://test.routledgehandbooks.com/doi/10.1201/b10440-5
K. Trivedi Pravin, K. Munkin Murat
Published online on: 20 Dec 2010

4

Recent Developments in Cross Section and Panel Count Models

Pravin K. Trivedi and Murat K. Munkin

CONTENTS

4.1 Introduction

Count data regression is now a well-established tool in econometrics. If the outcome variable is measured as a nonnegative count, y, $y \in \mathbb{N}_0 = \{0, 1, 2, \dots\}$, and the object of interest is the marginal impact of a change in the variable x on the regression function $\mathrm{E}[y|x]$, then a count regression is a relevant tool of analysis. Because the response variable is discrete, its distribution places probability mass at nonnegative integer values only. Fully parametric formulations of count models accommodate this property of the distribution. Some semiparametric regression models only accommodate $y \geq 0$, but not discreteness. Given the discrete nature of the outcome variable, a linear regression is usually not the most efficient method of analyzing such data. The standard count model is a nonlinear regression.

Several special features of count regression models are intimately connected to discreteness and nonlinearity. As in the case of binary outcome models like the logit and probit, the use of count data regression models is very widespread in empirical economics and other social sciences. Count regressions have been extensively used for analyzing event count data that are common in fertility analysis, health care utilization, accident modeling, insurance, recreational demand studies, analysis of patent data.

Cameron and Trivedi (1998), henceforth referred to as CT (1998), and Winkelmann (2005) provided monograph length surveys of econometric count data methods. More recently, Greene (2007b) has also provided a selective survey of newer developments. The present survey also concentrates on newer developments, covering both the probability models and the methods of estimating the parameters of these models, as well as noteworthy applications or extensions of older topics. We cover specification and estimation issues at greater length than testing.

Given the length restrictions that apply to this article, we will cover cross-section and panel count regression but not time series count data models. The reader interested in time series of counts is referred to two recent survey papers; see Jung, Kukuk, and Liesenfeld (2006), and Davis, Dunsmuir, and Streett (2003). A related topic covers hidden Markov models (multivariate

time series models for discrete data) that have been found very useful in modeling discrete time series data; see MacDonald and Zucchini (1997). This topic is also not covered even though it has connections with several themes that we do cover.

The natural stochastic model for counts is derived from the Poisson point process for the occurrence of the event of interest, which leads to Poisson distribution for the number of occurrences of the event, with probability mass function

$$\Pr[Y = y] = \frac{e^{-\mu}\mu^{y}}{y!}, \qquad y = 0, 1, 2, \ldots, \tag{4.1}$$

where μ is the intensity or rate parameter. The first two moments of this distribution, denoted $\mathcal{P}[\mu]$, are $\mathrm{E}[Y] = \mu$, and $\mathrm{V}[Y] = \mu$, demonstrating the well-known equidispersion property of the Poisson distribution. The Poisson regression follows from the parameterization $\mu = \mu(\mathbf{x})$, where $\mathbf{x}$ is a K-dimensional vector of exogenous regressors. The usual specification of the conditional mean is

$$\mathrm{E}[y|\mathbf{x}] = \exp(\mathbf{x}'\beta). \tag{4.2}$$

Standard estimation methods are fully parametric Poisson maximum likelihood, or "semiparametric" methods such as nonlinear least squares, or moment-based estimation, based on the moment condition $\mathrm{E}[y - \exp(\mathbf{x}'\beta)|\mathbf{x}] = \mathbf{0}$, possibly further augmented by the equidispersion restriction used to generate a weight function.

Even when the analysis is restricted to cross-section data with strictly exogenous regressors, the basic Poisson regression comes up short in empirical work in several respects. The mean-variance equality restriction is inconsistent with the presence of significant unobserved heterogeneity in cross-section data. This feature manifests itself in many different ways. For example, Poisson model often under-predicts the probability of zero counts, in a data situation often referred to as the *excess zeros* problem. A closely related deficiency of the Poisson is that in contrast to the equidispersion property, data more usually tend to be overdispersed, i.e., (conditional) variance usually exceeds the (conditional) mean. Overdispersion can result from many different sources (see CT, 1998, 97–106). Overdispersion can also lead to the problem of excess zeros (or *zero inflation*) in which there is a much larger probability mass at the zero value than is consistent with the Poisson distribution. The literature on new functional forms to handle overdispersion is already large and continues to grow. Despite the existence of a plethora of models for overdispersed data, a small class of models, including especially the negative binomial regression (NBR), the two-part model (TPM), and the zero-inflated Poisson (ZIP) and zero-inflated negative binomial (ZINB), has come to dominate the applied literature. In what follows we refer to this as the set of basic or benchmark parametric count regression models, previously comprehensively surveyed in CT (1998, 2005).

Beyond the cross-section count regression econometricians are also interested in applying count models to time series, panel data, as well as multivariate models. These types of data generally involve patterns of dependence more general than those for cross-section analysis. For example, serial dependence of outcomes is likely in time series and panel data, and a variety of dependence structures can arise for multivariate count data. Such data provide considerable opportunity for developing new models and methods.

Many of the newer developments surveyed here arise from relaxing the strong assumptions underlying the benchmark models. These new developments include the following:

- A richer class of models of unobserved heterogeneity some of which permit nonseparable heterogeneity
- A richer parameterization of regression functions
- Relaxing the assumption of conditional independence of $y_i|\mathbf{x}_i$ $(i = 1, \ldots, N)$
- Relaxing the assumption that the regressors $\mathbf{x}_i$ are exogenous
- Allowing for self-selection in the samples
- Extending the standard count regression to the multivariate case
- Using simulation-based estimation to handle the additional complications due to more flexible functional form assumptions

The remainder of the chapter is arranged as follows. Section 4.2 concentrates on extensions of the standard model involving newer functional forms. Section 4.3 deals with issues of cross-sectional dependence in count data. Section 4.4 deals with the twin interconnected issues of count models with endogenous regressors and/or self-selection. Sections 4.4 and 4.5 cover panel data and multivariate count models, respectively. The final Section 4.6 covers computational matters.

4.2 Beyond the Benchmark Models

One classic and long-established extension of the Poisson regression is the negative binomial (NB) regression. The NB distribution can be derived as a Poisson-Gamma mixture. Given the Poisson distribution $f(y|x, \nu) = \exp(-\mu\nu)(\mu\nu)^y/y!$ with the mean $\mathrm{E}[y|\mathbf{x}, \nu] = \mu(\mathbf{x})\nu$, $\nu > 0$, where the random variable ν, representing multiplicative unobserved heterogeneity, a latent variable, has Gamma density $g(\nu) = \nu^{\alpha-1}\exp(-\nu)/\Gamma(\alpha)$, with $\mathrm{E}[\nu] = 1$, and variance $\alpha(\alpha > 0)$. The resulting mixture distribution is the NB:

$$f(y|\mu(\mathbf{x})) = \int_0^\infty f(y|\mu(\mathbf{x}), \nu)g(\nu)d\nu = \frac{\mu(\mathbf{x})^y\Gamma(y+\alpha)}{y!\Gamma(\alpha)}\left(\frac{1}{\mu(\mathbf{x})+\alpha}\right)^{y+\alpha}, \quad (4.3)$$

which has mean $\mu(\mathbf{x})$ and variance $\mu(\mathbf{x})[\mathbf{1} + \alpha\mu(\mathbf{x})] > \mathrm{E}[y|\mathbf{x}]$, thus accommodating the commonly observed overdispersion. The gamma heterogeneity assumption is very convenient, but the same approach can be used with other mixing distributions.

This leading example imposes a particular mathematical structure on the model. Specifically, the latent variable reflecting unobserved heterogeneity is separable from the main object of identification, the conditional mean. This is a feature of many established mixture models. Modern approaches, however, deal with more flexible models where the latent variables are nonseparable. In such models unobserved heterogeneity impacts the entire disribution of the outcome of interest. Quantile regression and finite mixtures are two examples of such nonseparable models.

There are a number of distinctive ways of allowing for unobserved heterogeneity. It may be treated as an additive or a multiplicative random effect (uncorrelated with included regressors) or a fixed effect (potentially correlated with included regressors). Within the class of random effects models, heterogeneity distributions may be treated as continuous or discrete. Examples include a random intercept in cross-section and panel count models, fixed effects in panel models of counts. Second, both intercept and slope parameters may be specified to vary randomly and parametrically, as in finite mixture count models. Third, heterogeneity may be modeled in terms of both observed and unobserved variables using mixed models, hierarchical models and/or models of clustering. The approach one adopts and the manner in which it is combined with other assumptions has important implications for computation. The second and third approaches are reflected in many recent developments.

4.2.1 Parametric Mixtures

The family of random effects count models is extensive. In Table 4.1 we show some leading examples that have featured in empirical work. By far the most popular is the negative binomial specification with either a linear variance function (NB1) or a quadratic variance function (NB2). Both these functional forms capture extra-Poisson probability mass at zero and in the right tail, as would other mixtures, e.g., Poisson lognormal. But the continuing popularity of the NB family rests on computational convenience, even though (as we discuss later in this chapter) computational advances have made other models empirically accessible. When the right tail of the distribution is particularly heavy, the Poisson-inverse Gaussian mixture (P-IG) with a cubic variance function is attractive, but again this consideration must be balanced against additional computational complexity (see Guo and Trivedi 2002).

The foregoing models are examples of continuous mixture models based on a continuous distribution of heterogeneity. Mixture models that also allow for finite probability point mass, such as the hurdle ("two part") model and zero inflated models shown in Table 4.1, that appeared in the literature more than a decade ago (see Gurmu and Trivedi 1996) have an important advantage – they

TABLE 4.1

Selected Mixture Models

Distribution	$f(y) = \Pr[Y = y]$	Mean; Variance
1 Poisson	$e^{-\mu}\mu^y/y!$	$\mu(\mathbf{x})$; $\mu(\mathbf{x})$
2 NB1	As in NB2 below with α^{-1} replaced by $\alpha^{-1}\mu$	$\mu(\mathbf{x})$; $(1+\alpha)\,\mu(\mathbf{x})$
3 NB2	$\dfrac{\Gamma(\alpha^{-1}+y)}{\Gamma(\alpha^{-1})\Gamma(y+1)}\left(\dfrac{\alpha^{-1}}{\alpha^{-1}+\mu}\right)^{\frac{1}{\alpha}}\left(\dfrac{\mu}{\mu+\alpha^{-1}}\right)^{y}$	$\mu(\mathbf{x})$; $(1+\alpha\mu(\mathbf{x}))\,\mu(\mathbf{x})$
4 P-IG	$\Pr(Y=0)\times\dfrac{\mu^k}{\Gamma(k+1)}(1+2\eta)^{-k/2}$	$\mu(\mathbf{x})$; $\mu(\mathbf{x})+\mu(\mathbf{x})^3/\tau$
$k\geq 1$	$\times\sum_{i=0}^{k-1}\dfrac{\Gamma(k+i)}{\Gamma(k-i)\Gamma(i+1)}\left(\left(\dfrac{\eta}{2\mu(\mathbf{x})}\right)^{i}(1+2\eta)^{-i/2}\right.$, where $\Pr(Y=0)=\exp\left[\dfrac{\mu}{\eta}(1-\sqrt{1+2\eta})\right]$, $(\eta=\mu^2/\tau)$	
5 Hurdle	$\begin{cases} f_1(0) & \text{if } y=0,\\ \dfrac{1-f_1(0)}{1-f_2(0)}f_2(y) & \text{if } y\geq 1.\end{cases}$	$\Pr[y>0\|\mathbf{x}]\mathrm{E}_{y>0}[y\|y>0,\mathbf{x}]$; $\Pr[y>0\|\mathbf{x}]\mathrm{V}_{y>0}[y\|y>0,\mathbf{x}]$ $+\Pr[y=0\|\mathbf{x}]\mathrm{E}_{y>0}[y\|y>0\|\mathbf{x}]$
6 Zero-inflated	$\begin{cases} f_1(0)+(1-f_1(0))f_2(0) & \text{if } y=0,\\ (1-f_1(0))f_2(y) & \text{if } y\geq 1.\end{cases}$	$(1-f_1(0))(\mu(\mathbf{x})+f_1(0)\mu^2(\mathbf{x}))$
7 Finite mixture	$\sum_{j=1}^{m}\pi_j f_j(y\|\theta_j)$	$\Sigma_{i=1}^{2}\pi_i\mu_i(\mathbf{x})$; $\Sigma_{i=1}^{2}\pi_i[\mu_i(\mathbf{x})+\mu_i^2(\mathbf{x})]$
8 PPp	$h_2(y\|\mu,\mathbf{a})=\dfrac{e^{-\mu}\mu^y}{y!}\dfrac{(1+a_1y+a_2y^2)^2}{\eta_2(\mathbf{a},\mu)}$ where $\eta_2(\mathbf{a},\mu)=1+2a_1m_1+(a_1^2+2a_2)m_2+2a_1a_2m_3+a_2^2m_4$	Complicated

relax the restrictions on both the conditional mean and variance functions. There are numerous ways of attaining such an objective using latent variables, latent classes, and a combination of these. This point is well established in the literature on generalized linear models. Skrondal and Rabe-Hesketh (2004) is a recent survey.

4.2.1.1 Hurdle and Zero-Inflated Models

Hurdle and zero-inflated models are motivated by the presence of "excess zeros" in the data. The hurdle model or two-part model (TPM) relaxes the assumption that the zeros and the positives come from the same data-generating process. Suppressing regressors for notational simplicity, the zeros are determined by the density $f_1(\cdot)$, so that $\Pr[y = 0] = f_1(0)$ and $\Pr[y > 0] = 1 - f_1(0)$. The positive counts are generated by the truncated density $f_2(y|y > 0) = f_2(y)/(1 - f_2(0))$, that is multiplied by $\Pr[y > 0]$ to ensure a proper distribution. Thus, $f(y) = f_1(0)$ if $y = 0$ and $f(y) = [1 - f_1(0)]f_2(y)/[1 - f_2(0)]$ if $y \geq 1$. This generates the standard model only if $f_1(\cdot) = f_2(\cdot)$.

Like the hurdle model, zero-inflated model supplements a count density $f_2(\cdot)$ with a binary process with density $f_1(\cdot)$. If the binary process takes value 0, with probability $f_1(0)$, then $y = 0$. If the binary process takes value 1, with probability $f_1(1)$, then y takes count values $0, 1, 2, \ldots$ from the count density $f_2(\cdot)$. This lets zero counts occur in two ways: either as a realization of the binary process or a count process. The zero-inflated model has density $f(y) = f_1(0) + [1 - f_1(0)]f_2(0)$ if $y = 0$, and $f(y) = [1 - f_1(0)]f_2(y)$ if $y \geq 1$. As in the case of the hurdle model the probability $f_1(0)$ may be parameterized through a binomial model like the logit or probit, and the set of variables in the $f_1(\cdot)$ density may differ from those in the $f_2(\cdot)$ density.

4.2.1.1.1 Model Comparison in Hurdle and ZIP Models Zero-inflated variants of the Poisson (ZIP) and the negative binomial (ZINB) are especially popular. For the empirical researcher this generates an embarrassment of riches. The challenge comes from having to evaluate the goodness of fit of these models and selecting the "best" model according to some criterion, such as the AIC or BIC. It is especially helpful to have software that can simultaneously display the relevant information for making an informed choice. Care must be exercised in model selection because even when the models under comparison have similar overall fit, e.g., log-likelihood, they may have substantially different implications regarding the marginal effect parameters, i.e., $\partial \mathrm{E}[y|\mathbf{x}]/\partial x_j$. A practitioner needs suitable software for model interpretation and comparison.

A starting point in model selection is provided by a comparison of fitted probabilities of different models and the empirical frequency distribution of counts. Lack of fit at specific frequencies may be noticeable even in an informal comparison. Implementing a formal goodness-of-fit model comparison is easier when the rival models are nested, in which case we can apply a likelihood ratio test. However, some empirically interesting pairs of models are not

nested, e.g., Poisson and ZIP, and negative binomial and ZINB. In these cases the so-called Vuong test (Vuong 1989), essentially a generalization of the likelihood ratio test, may be used to test the null hypothesis of equality of two distributions, say f and g. For example, consider the log of the ratio of fitted probabilities of Poisson and ZIP models, denoted $r_i = \ln\{\widehat{\Pr}_P(y_i|\mathbf{x}_i)/\widehat{\Pr}_{ZIP}(y_i|\mathbf{x}_i)\}$. Let $\overline{r} = N^{-1}\sum r_i$ and s_r denotes the standard deviation of r_i; then the test statistic $T_{vuong} = \overline{r}/(s_r/\sqrt{N})$ has asymptotic standard normal distribution. So the test can be based on the critical values of the standard normal. A large value of T_{vuong} in this case implies a departure from the null in the direction of Poisson, and a large negative value in the direction of ZIP. For other empirically interesting model pairs, e.g., ZIP and ZINB, the same approach can be applied, although it is less common for standard software to make these statistics available also. In such cases model selection information criteria such as the AIC and BIC are commonly used.

Two recent software developments have been very helpful in this regard. First, these models are easily estimated and compared in many widely used microeconometrics packages such as Stata and Limdep; see, for example, CT (2009) and Long and Freese (2006) for coverage of options available in Stata. For example, Stata provides goodness-of-fit and model comparison statistics in a convenient tabular form for the Poisson, NB2, ZIP, and ZINB. Using packaged commands it has become easy to compare the fitted and empirical frequency distribution of counts in a variety of parametric models. Second, mere examination of estimated coefficients and their statistical significance provides an incomplete picture of the properties of the model. In empirical work, a key parameter of interest is the average marginal effect (AME), $N^{-1}\sum_{i=1}^{N}\partial \mathrm{E}[y_i|\mathbf{x}_i]/\partial x_{j,i}$, or the marginal effect evaluated at a "representative" value of $\mathbf{x}$ (MER). Again, software developments have made estimation of these parameters very accessible.

4.2.1.2 Finite Mixture Specification

An idea that is not "recent" in principle, but has found much traction in recent empirical work of discrete or mixtures of count distributions. Unlike the NB model, which has a continuous mixture representation, the finite mixture approach instead assumes a discrete representation of unobserved heterogeneity. It encompasses both intercept and slope heterogeneity and hence the full distribution of outcomes. This generates a class of flexible parametric models called finite mixture models (FMM) – a subclass of latent class models; see Deb (2007), CT (2005, Chapter 20.4.3).

A FMM specifies that the density of y is a linear combination of m different densities, where the jth density is $f_j(y|\beta_j)$, $j = 1, 2, \ldots, m$. An m-component finite mixture is defined by

$$f(y|\beta, \pi) = \sum_{j=1}^{m} \pi_j f_j(y|\beta_j), \quad 0 < \pi_j < 1, \sum_{j=1}^{m} \pi_j = 1. \tag{4.4}$$

A simple example is a two-component ($m = 2$) Poisson mixture of $\mathcal{P}[\mu_1]$ and $\mathcal{P}[\mu_2]$. This may reflect the possibility that the sampled population contains

two "types" of cases, whose y outcomes are characterized by distributions $f_1(y|\beta_1)$ and $f_2(y|\beta_2)$ that are assumed to have different moments. The mixing fraction π_1 is in general an unknown parameter. In a more general formulation it too can be parameterized in terms of observed variable(s) $\mathbf{z}$.

The FMM specification is attractive for empirical work in cross-section analysis because it is flexible. Mixture components may come from different parametric families, although commonly they are specified to come from the same family. The mixture components permit differences in conditional moments of the components, and hence in the marginal effects. In an actual empirical setting, the latent classes often have a convenient interpretation in terms of the differences between the underlying subpopulations.

Application of FMM to panel data is straightforward if the panel data can be treated as pooled cross section. However, when the T-dimension of a panel is high in the relevant sense, a model with fixed mixing probabilities may be tenuous as transitions between latent classes may occur over time. Endogenous switching models allow the transition probability between latent classes to be correlated with outcomes and hidden Markov models allow the transition probabilities to depend upon past states; see Fruhwirth-Schnatter (2006) and MacDonald and Zucchini (1997).

There are a number of applications of the FMM framework for cross-section data. Deb and Trivedi (1997) use Medical Expenditure Panel Survey data to study the demand for care by the elderly using models of two- and three-component mixtures of several count distributions. Deb and Trivedi (2002) re-examine the Rand Health Insurance Experiment (RHIE) pooled cross-section data and show that FMM fit the data better than the hurdle (two-part) model. Of course, this conclusion, though not surprising, is specific to their data set. Lourenco and Ferreira (2005) apply the finite mixture model to model doctor visits to public health centers in Portugal using truncated-at-zero samples. Bohning and Kuhnert (2006) study the relationship between mixtures of truncated count distributions and truncated mixture distributions and give conditions for their equivalence.

Despite its attractions, the FMM class has potential limitations. First, maximum likelihood (ML) estimation is not straightforward because, in general, the log-likelihood function may have multiple maxima. The difficulties are greater if the mixture components are not well separated. Choosing a suitable optimization algorithm is important. Second, it is easy to overparameterize mixture models. When the number of components is small, say 2, and the means of the component distribution are far apart, discrimination between the components is easier. However, as additional components are added, there is a tendency to "split the difference" and unambiguous identification of all components becomes difficult because of the increasing overlap in the distributions. In particular, the presence of outliers may give rise to components that account for a small proportion (small values of π_j) of the observations. That is, identification of individual components may be fragile. CT (2009, Chapter 17) give examples using Stata's FMM estimation (Deb, 2007) command and suggest practical ways of detecting estimation problems.

Recent biometric literature offers promise of more robust estimation of finite mixtures via alternative to maximum likelihood. Lu, Hui, and Lee (2003), following Karlis and Xekalaki (1998), use minimum Hellinger distance estimation (MHDE) for finite mixtures of Poisson regressions; Xiang et al. (2008) use MHDE for estimating a k-component Poisson regression with random effects. The attraction of MHDE relative to MLE is that it is expected to be more robust to the presence of outliers and when mixture components are not well separated, and/or when the model fit is poor.

4.2.1.3 Hierarchical Models

While cross-section and panel data are by far the most common in empirical econometrics, sometimes other data structures are also available. For example, sample survey data may be collected using a multi-level design; an example is state-level data further broken down by counties, or province-level data clustered by communes (see Chang and Trivedi [2003]). When multi-level covariate information is available, hierarchical modeling becomes feasible. Such models have been widely applied to the generalized linear mixed model (GLMM) class of which Poisson regression is a member. For example, Wang, Yau, and Lee (2002) consider a hierarchical Poisson mixture regression to account for the inherent correlation of outcomes of patients clustered within hospitals. In their set-up data are in m clusters, with each cluster having n_j ($j = 1, \ldots, m$) observations, let $n = \sum n_j$. For example, the following Poisson-lognormal mixture can be interpreted as a one-level hierarchical model.

$$\begin{aligned} y_{ij} &\sim \mathcal{P}(\mu_{ij}), \qquad i = 1, \ldots, n_j;\ j = 1, \ldots, m \\ \log \mu_{ij} &= \mathbf{x}_{ij}'\beta + \varepsilon_{ij},\ \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2). \end{aligned} \tag{4.5}$$

An example of a two-level model, also known as a hierarchical Poisson mixture, that incorporates covariate information at both levels is as follows:

$$y_{ij} \sim \mathcal{P}(\mu_{ij}), \qquad i = 1, \ldots, n_j;\ j = 1, \ldots, m \tag{4.6}$$

$$\begin{aligned} \log \mu_{ij} &= \mathbf{x}_{ij}'\beta_j + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \mathcal{N}\left(0, \sigma_\varepsilon^2\right) \\ \beta_{kj} &= \mathbf{w}_{kj}'\gamma + v_{kj}; \quad v_{kj} \sim \mathcal{N}\left(0, \sigma_v^2\right),\ k = 1, \ldots K;\ j = 1, \ldots., m. \end{aligned} \tag{4.7}$$

In this case coefficients vary by clusters, and cluster-specific variables $\mathbf{w}_{kj}$ enter at the second level to determine the first-level parameters β_j, whose elements are β_{kj}. The parameter vector γ, also called hyperparameter, is the target of statistical inference. Both classical (Wang, Yau, and Lee 2002) and Bayesian analyses can be applied.

4.2.2 Quantile Regression for Counts

Quantile regression (QR) is usually applied to continuous response data; see Koenker (2005) for a thorough treatment of properties of QR. QR is consistent under weak stochastic assumptions and is equivariant to monotone

transformations. A major attraction of QR is that it potentially allows for response heterogeneity at different conditional quantiles of the variables of interest. If the method could be extended to counts, then one could go beyond the standard and somewhat restrictive models of unobserved heterogeneity based on strong distributional assumptions. Also QR facilitates a richer interpretation of the data because it permits the study of the impact of regressors on both the location and scale parameters of the model, while at the same time avoiding strong distributional assumptions about data. Moreover, advances made in quantile regression such as handling endogenous regressors can be exploited for count data. The problem, however, is that the quantiles of discrete variables are not unique since the c.d.f. is discontinuous with discrete jumps between flat sections. By convention the lower boundary of the interval defines the quantile in such a case. However, recent theoretical advances have extended QR to a special case of count regression; see Machado and Santos Silva (2005), Miranda (2006, 2008), Winkelmann (2006).

The key step in the quantile count regression (QCR) model of Machado and Santos Silva (2005) involves replacing the discrete count outcome y with a continuous variable $z = h(y)$, where $h(\cdot)$ is a smooth continuous transformation. The standard linear QR methods are then applied to z. The particular continuation transformation used is $z = y + u$, where $u \sim \mathcal{U}[0, 1]$ is a pseudo-random draw from the uniform distribution on $(0, 1)$. This step is called "jittering" the count. Point and interval estimates are then retransformed to the original y-scale using functions that preserve the quantile properties.

Let $Q_q(y|\mathbf{x})$ and $Q_q(z|\mathbf{x})$ denote the qth quantiles of the conditional distributions of y and z, respectively. The conditional quantile for $Q_q(z|\mathbf{x})$ is specified to be

$$Q_q(z|\mathbf{x}) = q + \exp(\mathbf{x}'\beta_q). \tag{4.8}$$

The additional term q appears in the equation because $Q_q(z|\mathbf{x})$ is bounded from below by q, due to the jittering operation.

To be able to estimate a quantile model in the usual linear form $\mathbf{x}'\beta$, a log transformation is applied so that $\ln(z - q)$ is modelled, with the adjustment that if $z - q < 0$ then we use $\ln(\varepsilon)$ where ε is a small positive number. The transformation is justified by the equivariance property of the quantiles and the property that quantiles above the censoring point are not affected by censoring from below. Post-estimation transformation of the z-quantiles back to y-quantiles uses the ceiling function, with

$$Q_q(y|\mathbf{x}) = \lceil Q_q(z|\mathbf{x}) - 1 \rceil, \tag{4.9}$$

where the symbol $\lceil r \rceil$ in the right-hand side of Equation 4.9 denotes the smallest integer greater than or equal to r.

To reduce the effect of noise due to jittering, the model is estimated multiple times using independent draws from $\mathcal{U}(0, 1)$ distribution, and the multiple estimated coefficients and confidence interval endpoints are averaged.

Hence the estimates of the quantiles of y counts are based on $\widehat{Q}_q(y|\mathbf{x}) = \lceil Q_q(z|\mathbf{x}) - 1 \rceil = \lceil q + \exp(\mathbf{x}'\overline{\widehat{\beta}}_q) - 1 \rceil$, where $\overline{\widehat{\beta}}$ denotes the average over the jittered replications.

Miranda (2008) applies the QCR to analysis of Mexican fertility data. Miranda (2006) describes Stata's add-on `qcount` command for implementing QCR. CT (2009, Chapter 7.5) discuss an empirical illustration in detail, with special focus on marginal effects. The specific issue of how to choose the quantiles is discussed by Winkelmann (2006), the usual practice being to select a few values such as q equal to 25, .50, and .75. This practice has to be modified to take account of the zeros problem because it is not unusual to have (say) 35% zeros in a sample, in which case q must be greater than .35.

4.3 Adjusting for Cross-Sectional Dependence

The assumption of cross-sectionally independent observations was common in the econometric count data literature during and before the 1990s. Recent theoretical and empirical work pays greater attention to the possibility of cross-sectional dependence. Two sources of dependence in cross-sectional data are stratified survey sampling design and, in geographical data, dependence due spatially correlated unobserved variables.

Contrary to a common assumption in cross-section regression, count data used in empirical studies are more likely to come from complex surveys derived from stratified sampling. Data from the stratified random survey samples, also known as complex surveys, are usually dependent. This may be due to use of survey design involving interviews with multiple households in the same street or block that may be regarded as natural clusters, where by cluster is meant a set whose elements are subject to common shocks. Such a sampling scheme is likely to generate correlation within cluster due to variation induced by common unobserved cluster-specific factors. Cross-sectional dependence between outcomes invalidates the use of variance formulae based on assumption of simple random samples.

Cross-sectional dependence also arises when the count outcomes have a spatial dimension, as when the data are drawn from geographical regions. In such cases the outcomes of units that are spatially contiguous may display dependence that must be controlled for in regression analysis.

There are two broad approaches for controlling for dependence within cluster, the key distinction being between random and fixed cluster effects analogous to panel data analysis.

4.3.1 Random Effects Cluster Poisson Regression

To clarify this point additional notation is required. Consider a sample with total N observations, which are distributed in C clusters with each cluster

having $N_c (c = 1, \ldots, C)$ observations and $\sum_{c=1}^{C} N_c = N$. If the number of observations per cluster varies, the data correspond to an unbalanced panel. Intra-cluster correlation refers to correlation between $y_{i,c}$ and $y_{j,c}$, $i \neq j$, $c = 1, \ldots, C$. A common assumption is of nonzero intra-cluster correlation and zero between-cluster correlation, i.e., corr $[y_{i,c}, y_{j,c'}, i \neq j, c \neq c'] = 0$. Additional complications arise according to the assumptions regarding N_c and C, i.e., whether there are many small clusters or few large clusters. The notation for handling clusters is similar to that for panel data, and a number of important results we cover also parallel similar ones in the panel data literature.

For specificity, we consider the Poisson regression for clustered data. A popular assumption states that the cluster-specific effects enter the model through the intercept term alone. The clustered count data, denoted $y_{ij}, i = 1, \ldots, C$, $j = 1, \ldots, N_j$, are Poisson distributed with $\mathrm{E}[y_{ij}|\mathbf{x}_{ij}, \alpha_i] = \exp(\alpha_i + \alpha + \mathbf{x}'_{ij}\beta)$, where $\mathbf{x}_{ij}$ are linearly independent covariates; the term α_i is the deviation of cluster-specific intercept from the population-averaged fixed intercept α. This model is referred to as cluster-specific intercept Poisson regression. A number of results are available corresponding to different assumptions about α_i. Demidenko (2007) presents several results for the case in which $\alpha_i (i = 1, \ldots, C)$ are i.i.d., C approaches ∞, and $N_c < \infty$. In the econometrics literature this corresponds to the cluster-specific random effects (CSRE) Poisson regression.

1. Under the assumptions stated above, the standard M-estimators for the Poisson regression applied to pooled clustered data are consistent but not efficient.
2. In the case where C is relatively small, e.g., $C < \min N_i$, a separate dummy variable corresponding to each cluster-specific intercept can be introduced in the conditional mean function and standard maximum likelihood procedure can be applied to the resulting model.
3. Under a specific assumption about the conditional covariance structure for the data, a generalized estimating equations (GEE), or (in econometrics terminology) nonlinear generalized least squares, procedure may be applied for efficiency gain over the simple Poisson. This requires a working matrix as an estimator of the unknown true variance matrix. Under the assumption that the cluster-random effects are equicorrelated, the working matrix can be parameterized in terms of a single correlation parameter.
4. Under a strong parametric assumption about the distribution of α_i, maximum likelihood can be applied. Computationally this is more demanding and may require simulation-based estimation.
5. Under somewhat special assumptions in which all clusters have the same number of observations, and the covariates are identical across clusters the methods mentioned above are all equivalent.

4.3.2 Cluster-Robust Variance Estimation

An approach that does not require a distributional assumption for the random component is to simply use robust variance estimation, i.e., "cluster robust" standard errors obtained by adapting the so-called Eicker–White robust variance estimator to handle clustered data. Specifically, if f_i ($i = 1, \ldots, n$) denotes the density for the ith observation, θ denotes the vector of unknown parameters, then the cluster-robust variance estimator evaluated at the maximum likelihood estimate $\widehat{\theta}_{MLE}$ is given by

$$V_C = \left[\sum_{j=1}^{C}\sum_{i=1}^{N_j}\frac{\partial^2 \ln f_{ij}}{\partial\theta\partial\theta'}\right]^{-1}\left[\sum_{j=1}^{C}\sum_{i=1}^{N_j}\sum_{k=1}^{N_j}\frac{\partial \ln f_{ij}}{\partial\theta}\frac{\partial \ln f_{kj}}{\partial\theta'}\right] \times \left[\sum_{j=1}^{C}\sum_{i=1}^{N_j}\frac{\partial^2 \ln f_{ij}}{\partial\theta\partial\theta'}\right]^{-1}\Bigg|_{\widehat{\theta}_{MLE}}. \tag{4.10}$$

If, instead of ML estimation, another consistent M-estimator is used, e.g., that defined by the nonlinear estimating equations $\sum_{c=1}^{C}\sum_{j=1}^{N_c}\mathbf{h}(y_{jc}, \mathbf{x}_{jc}, \theta) = \mathbf{0}$, the above formula is adjusted by replacing the score function $\partial \ln f_{ij}/\partial\theta$ by $\mathbf{h}_{i,j}(\widehat{\theta})$; see CT (2005, Chapter 24.5.6). Observe how within each cluster we do not use the likelihood score for each observation as in the case of independent observations; instead we replace it by the sum of likelihood scores over all cluster elements. The usual regularity conditions for the validity of the "sandwich" variance formula are required.

4.3.3 Cluster-Specific Fixed Effects

The Poisson fixed effects cluster model specifies

$$\begin{aligned} &y_{ij} \sim \mathcal{P}[\mu_{ij}], \quad i = 1, \ldots, C,\ j = 1, \ldots, N_i, \\ &\mu_{ij} = \alpha_i \exp(\mathbf{x}_{ij}'\beta), \end{aligned} \tag{4.11}$$

where $\mathbf{x}_{ij}$ excludes an intercept and any cluster-invariant regressors. The difference from the standard Poisson model is that the usual conditional mean $\exp(\mathbf{x}_{ij}'\beta)$ is scaled multiplicatively by the cluster-specific fixed effect (FE) α_i. Because of its similarity to the panel Poisson model, we defer a longer treatment of estimation to Section 4.5, but simply note that one can use either the conditional maximum likelihood approach in which inference is carried out conditionally on sufficient statistics for the fixed effects (i.e., the parameters α_i are eliminated), or we can introduce cluster-specific dummy variables and apply the standard ML estimation.

4.3.4 Spatial Dependence

We now consider models in which outcomes are counts with a spatial distribution; hence it becomes necessary to adjust for spatial correlation between

neighboring counts. Such spatial dependence is characterized by some underlying data generating process. Griffith and Haining (2006) survey the early literature on spatial Poisson regression and a number of its modern extensions.

Besag (1974) defined a general class of "auto-models" suitable for different types of spatially correlated data. In the special case, dubbed the "auto-Poisson" model, $\mathcal{P}[Y(i) = y(i)|\{Y(j)\},\ j \in \mathsf{N}(i)]$ denotes the conditional probability that the random variable $Y(i)$, defined at location i, realizes value $y(i)$, given the values of Y at the sites in the neighborhood of i, denoted $\mathsf{N}(i)$. If the $\{Y(i)\}$ have an auto-Poisson distribution with intensity parameter $\mu(i)$, then

$$\mathcal{P}[Y(i) = y(i)|\{Y(j)\},\ j \in \mathsf{N}(i)] = \frac{e^{-\mu(i)}\mu(i)^{y(i)}}{y(i)!}$$

$$\log \mu(i) = \alpha(i) + \sum_{j \in \mathsf{N}(i)} \beta(i,\ j)y(j)) \qquad (4.12)$$

where the parameter $\alpha(i)$ is an area-specific effect, and $\beta(i,\ j) = \beta(j,\ i)$. The standard set-up specified $\beta(i,\ j) = \gamma w(i,\ j)$, where γ is a spatial autoregressive parameter, and $w(i,\ j)$ (= 0 or 1) represents the neighborhood structure. Let $\mathsf{N}(i)$ denote the set of neighbors of area i; then $w(i,\ j) = 1$ if i and j are neighbors $[j \in \mathsf{N}(i)]$, and otherwise $w(i,\ j) = 0$. In addition, $w(i, i) = 0$ must be assumed. The difficulty with this auto-Poisson model is under the restriction $\sum_i \mathcal{P}[Y(i) = y(i)|\{Y(j)\},\ j \in \mathsf{N}(i)] = 1$, implies $\gamma \leq 0$, which implies dependence property with negative spatial correlation. The spatial count literature has evolved in different directions to overcome this difficulty; see Kaiser and Cressie (1997) and Griffith and Haining (2006). One line of development uses spatial weighting dependent on assumptions about the spatial dependence structure.

A different approach for modeling either positive or negative spatial autocorrelation in a Poisson model is based on a mixture specification. The conditional mean function is defined as

$$\ln[\mu(i)] = \alpha(i) + S(i) \qquad (4.13)$$

where $S(i)$ is a random effect defined by a conditional autoregressive model with a general covariance structure. For example, an N-dimensional multivariate normal specification with covariance matrix $\sigma^2(\mathbf{I}_N - \tau\mathbf{W})^{-1}$, where $\mathbf{W} = [\mathbf{w}(i,\ j)]$ is the spatial weighting matrix, can capture negative or positive dependence between outcomes through the spatial autoregressive parameter τ. The resulting model is essentially a Poisson mixture model with a particular dependence structure. Estimation of this mixture specification by maximum likelihood would typically require simulation because its likelihood will not be expressed in a closed form. Griffith and Haining (2006) suggest Bayesian Markov chain Monte Carlo (MCMC); the use of this method is illustrated in Subsection 4.6.3. They compare the performance of several

alternative estimators for the Poisson regression applied to georeferenced data.

Spatial dependence modeled using a mixture specification induces overdispersion. If the conditional mean is correctly specified, the resulting model can be consistently estimated using the Poisson model under pseudo-likelihood. Robust variance estimator should be used to adjust for the effects of overdispersion. However, an advantage of estimating the full specification is that one obtains information about the structure of spatial dependence.

4.4 Endogeneity and Self-Selection

Endogenous regressors, both categorical and continuous, arise naturally in many count regression models. A well-known example from health economics involves models of counts of health services, e.g., doctor visits, with one of the regressors being the health insurance status of the individual. Assumption that choice of health insurance and the count outcome equation are conditionally uncorrelated is unrealistic when data are observational and insurance status is not exogenously assigned, rather it is self-selected. The case of endogenous dummy variables occurs commonly in empirical work and thus will get special attention in this section.

Important earlier analyses of count models with endogenous regressors include Mullahy (1997) and Windmeier and Santos Silva (1997) who proposed moment-based estimators within a GMM framework, and Terza (1998) who provided a full-information parametric analysis as well as a "semiparametric" sequential two-step estimator. They are motivated by a desire for more robust estimators. Discreteness of count outcomes, and often also that of the endogenous regressor, is typically ignored.

4.4.1 Moment-Based Estimation

Consider the exponential mean model $\mathrm{E}[y_i|\mathbf{x}_i] = \exp(\mathbf{x}_i'\beta)$, where at least one component of $\mathbf{x}$ is endogenous. To introduce endogeneity into this model the first step is to introduce another source of randomness in the specification. This is done using unobserved heterogeneity, either additively or multiplicatively. For example, Mullahy specified the moment as $\mathrm{E}[y_i|\mathbf{x}_i, \nu_i] = \exp(\mathbf{x}_i'\beta)\nu_i$, where $\mathrm{E}[\nu_i] = 1$, with $(\mathbf{x}_i, \nu_i)$ being jointly dependent. Instead of introducing further parametric assumptions about the dependence structure of $(\mathbf{x}_i, \nu_i)$, the GMM approach postulates the availability of a vector of instruments, $\mathbf{z}$, $\dim(\mathbf{z}) \geq \dim(\mathbf{x})$, that satisfy the moment condition

$$\mathrm{E}[y_i \exp(-\mathbf{x}_i'\beta) - v_i|\mathbf{z}_i] = \mathbf{0}. \tag{4.14}$$

By assumption, $\mathbf{z}$ is orthogonal to ν_i, and may have common elements with $\mathbf{x}$, but has some distinct elements that are excluded from $\mathbf{x}$. The instruments

are assumed to be valid and relevant, in the sense that there is nontrivial correlation between **z** and **x**. Note that this moment condition is different from that for the case in which ν_i enters the conditional mean additively – IVs that are orthogonal under multiplicative error are not so in general under additive errors.

A specific feature of the count model is discreteness and heteroskedasticity. GMM estimation of this model usually ignores the first feature; however, two-step efficient GMM estimators can accommodate heteroskedasticity. This topic has been surveyed in CT (2005) who also provide a discussion of the practical aspects of implementing efficient GMM estimation, but this discussion is broader in scope and not just for count regressions.

4.4.2 Control Function Approach

Since endogenous regressors cause significant complication in estimation of nonlinear models, one strategy is to first test for the presence of endogenous regressors, and then to use a suitable new estimator only if the null hypothesis of zero correlation between the regressors and the equation error is rejected. When the estimator is defined by a moment condition, the equation error is also implicitly defined by it. The error term, say u, is explicitly defined in linear two-stage least squares. There, and in the related literature on Durbin–Wu–Hausman tests of endogeneity (see Davidson and MacKinnon 2004, Chapter 8.7), the following test procedure is recommended. Suppose the regression of interest with dependent variable y_1 has a scalar right-hand side variables y_2 and exogenous variables $\mathbf{X}$. Let $\mathbf{W} = [\mathbf{Z}\ \mathbf{X}]$ denote the set of instrumental variables. Let $\mathbf{P}_W y_2$ be the linear projection of y_2 on $\mathbf{W}$. Then a test of endogeneity of y_2 is a test of $H_0 : \delta = 0$ in the OLS regression $y_1 = \mathbf{X}\beta_1 + y_2\beta_2 + \mathbf{P}_W y_2\delta + u$. Because of the least squares identity $y_2 \equiv P_W y_2 + \widehat{v}_2$, this procedure is equivalent to testing $H_0 : \delta^* = 0$ in the regression

$$y_1 = \mathbf{X}\beta_1 + (\beta_2 + \delta^*)y_2 - \delta^*\widehat{v} + u,$$

in which the right-hand side is augmented by the reduced form residual $\widehat{v}$. If the null hypothesis is rejected, this OLS regression is equivalent to the standard two-stage least squares. In essence, adding the variable $\widehat{v}_2$ controls for the endogeneity of y_2; once it is included the standard OLS estimator yields consistent point estimates. Hence we refer to this approach as the control function approach.

An interesting question is whether this approach can be extended to standard count regression models. Differences from the two-stage least squares case are due to the exponential conditional mean function and multiplicative error term. Terza (1998) considered a bivariate model in which the counted variable y depends on exogenous variables **x** and an endogenous treatment dummy variable $d(\equiv y_2)$. He provided both maximum likelihood and a two-step ("semiparametric") estimator for this model. His two-step estimator can be interpreted as a control function estimator.

Hardin, Schmiediche, and Carroll (2003) propose a method for estimating the Poisson regression with endogenous regressors that can also be interpreted as a control function type approach. Their method is intended to apply to models in the linear exponential family, Poisson regression being a special case. A linear-reduced form regression is estimated for the endogenous variable. As in two-stage least squares, valid instruments, $\mathbf{x}_2$, are assumed to be available. The linear-reduced form regression, estimated by OLS, generates predicted values for the endogenous variable y_2, denoted $\widehat{y}_2$. The original Poisson regression is estimated after replacing the endogenous variable by its predicted value. The variances are obtained using a bootstrap to allow for the fact that $\widehat{y}_2$ is a generated regressor subject to sampling variability.

Formally, the Poisson regression is estimated given the conditional mean function

$$E(y|\widehat{y}_2, x_1) = \mu = \exp(\beta_1\widehat{y}_2 + \mathbf{x}_1'\beta_2), \tag{4.15}$$

where $\widehat{y}_2 = \mathbf{x}_1'\widehat{\gamma}_1 + \mathbf{x}_2'\widehat{\gamma}_2$. This approach does not combine testing for endogeneity with estimation; it instead estimates the model assuming endogeneity. Despite its similarity with moment-based methods, the basis of the moment condition being used is not clear.

Finally we consider another somewhat ad hoc fitted-value method that resembles two-stage least squares that has been used in the context of Poisson regression with one endogenous dummy variable. This set-up is common in empirical work. Consider the overdispersed Poisson model,

$$\begin{aligned} y_i &\sim \mathcal{P}[\mu_i\eta_i], i = 1, \ldots, N \\ \mathrm{E}[y_i|\mu_i, \eta_i] &= \mu_i\eta_i \\ &= \exp(\mathbf{x}_i'\beta + \gamma d_i + \varepsilon_i), \end{aligned} \tag{4.16}$$

where $\eta_i = \exp(\varepsilon_i)$, d_i is the endogenous dummy variable, ε_i is unobserved heterogeneity uncorrelated with $\mathbf{x}_i'$.

Suppose $\mathbf{z}_i$ be a set of valid instruments, $\dim(\mathbf{z}_i) > \dim(\mathbf{x}_i)$. Assume $\mathrm{E}[y_i - \mu_i|\mathbf{z}_i] = \mathbf{0}$, but $\mathrm{E}[y_i - \mu_i|\mathbf{x}_i] \neq \mathbf{0}$. Consider the following two-step estimator: (1) generate a fitted value $\widehat{d}_i(\mathbf{z}_i)$ from a "reduced form" of d_i, using instruments $\mathbf{z}_i$; (2) replace d_i by $\widehat{d}_i(\mathbf{z}_i)$ and estimate the new Poisson regression by MLE. Though appealing in its logic, it is not clear that the two-step estimator is consistent. Let $d_i = \widehat{d}_i(\mathbf{z}_i) + \widehat{\nu}_i$, where $\widehat{d}_i(\mathbf{z}_i)$ is a predicted ("fitted") value of d_i from linear regression on $\mathbf{z}_i$. Hence,

$$\begin{aligned} \mathrm{E}(y_i|\widehat{d}_i, \widehat{\nu}_i) &= \exp(\mathbf{x}_i'\beta + \gamma\widehat{d}_i(\mathbf{z}_i) + \varepsilon_i + \gamma\widehat{\nu}_i) \\ &= \exp(\mathbf{x}_i'\beta + \gamma\widehat{d}_i(\mathbf{z}_i)) \exp(\varepsilon_i + \gamma\widehat{\nu}_i). \end{aligned} \tag{4.17}$$

Consistency requires $\mathrm{E}[\exp(\varepsilon_i + \gamma\widehat{\nu}_i)|\mathbf{x}_i, \widehat{d}_i] = \mathbf{0}$, but it is not obvious that this condition can be satisfied regardless of the functional form used to generate $\widehat{d}_i(\mathbf{z}_i)$.

4.4.3 Latent Factor Models

An alternative to the above moment-based approaches is a pseudo-FIML approach of Deb and Trivedi (2006a) who consider models with count outcome and endogenous treatment dummies. The model is used to study the impact of health insurance status on utilization of care. Endogeneity in these models arises from the presence of common latent factors that impact both the choice of treatments a (interpreted as treatment variables) and the intensity of utilization (interpreted as an outcome variable). The specification is consistent with selection on unobserved (latent) heterogeneity. In this model the endogenous variables in the count outcome equations are categorical, but the approach can be extended to the case of continuous variables.

The model includes a set of J dichotomous treatment variables that correspond to insurance plan dummies. These are endogenously determined by mixed multinomial logit structure (MMNL)

$$\Pr(\mathbf{d}_i|\mathbf{z}_i, \mathbf{l}_i) = \frac{\exp(\mathbf{z}_i'\alpha_j + \delta_j l_{ij})}{1 + \sum_{k=1}^{J} \exp(\mathbf{z}_i'\alpha_k + \delta_k l_{ik})}. \tag{4.18}$$

where d_j is observed treatment dummies, $\mathbf{d}_i = [d_{i1}, d_{i2}, \ldots, d_{iJ}]$, $j = 0, 1, 2, \ldots, J$, $\mathbf{z}_i$ is exogenous covariates, $\mathbf{l}_i = [l_{i1}, l_{i2}, \ldots, l_{iJ}]$, and l_{ij} are latent or unobserved factors.

The expected outcome equation for the counted outcomes is

$$\mathrm{E}(y_i|\mathbf{d}_i, \mathbf{x}_i, \mathbf{l}_i) = \exp\left(\mathbf{x}_i'\beta + \sum\nolimits_{j=1}^{J} \gamma_j d_{ij} + \sum\nolimits_{j=1}^{J} \lambda_j l_{ij}\right), \tag{4.19}$$

where $\mathbf{x}_i$ is a set of exogenous covariates. When the factor loading parameter $\lambda_j > 0$, treatment and outcome are positively correlated through unobserved characteristics, i.e., there is positive selection. Deb and Trivedi (2006a) assume that the distribution of y_i is negative binomial

$$f(y_i|\mathbf{d}_i, \mathbf{x}_i, \mathbf{l}_i) = \frac{\Gamma(y_i + \psi)}{\Gamma(\psi)\Gamma(y_i + 1)} \left(\frac{\psi}{\mu_i + \psi}\right)^{\psi} \left(\frac{\mu_i}{\mu_i + \psi}\right)^{y_i}, \tag{4.20}$$

where $\mu_i = \mathrm{E}(y_i|\mathbf{d}_i, \mathbf{x}_i, \mathbf{l}_i) = \exp(\mathbf{x}_i'\beta + \mathbf{d}_i'\gamma + \mathbf{l}_i'\lambda)$ and $\psi \equiv 1/\alpha$ ($\alpha > 0$) is the overdispersion parameter.

The parameters in the MMNL are only identified up to a scale. Hence a scale normalization for the latent factors is required; accordingly, they set $\delta_j = 1$ for each j. Although the model is identified through nonlinearity when $\mathbf{z}_i = \mathbf{x}_i$, they include some variables in $\mathbf{z}_i$ that are not included $\mathbf{x}_i$.

Joint distribution of treatment and outcome variables is

$$\begin{aligned} \Pr(y_i, \mathbf{d}_i|\mathbf{x}_i, \mathbf{z}_i, \mathbf{l}_i) &= f(y_i|\mathbf{d}_i, \mathbf{x}_i, \mathbf{l}_i) \times \Pr(\mathbf{d}_i|\mathbf{z}_i, \mathbf{l}_i) \\ &= f(\mathbf{x}_i'\beta + \mathbf{d}_i'\gamma + \mathbf{l}_i'\lambda) \\ &\quad \times \mathbf{g}(\mathbf{z}_i'\alpha_1 + \delta_1 l_{i1}, \ldots, \mathbf{z}_i'\alpha_J + \delta_J l_{iJ}). \end{aligned} \tag{4.21}$$

This model does not have a closed-form log-likelihood, but it can be estimated by numerical integration and simulation-based methods (Gourieroux

and Monfort 1997). Specifically, as l_{ij} are unknown, it is assumed that the l_{ij} are i.i.d. draws from (standard normal) distribution and one can numerically integrate over them.

$$\begin{aligned}\Pr(y_i, \mathbf{d}_i|\mathbf{x}_i, \mathbf{z}_i) &= \int \big[f(\mathbf{x}_i'\beta + \mathbf{d}_i'\gamma + \mathbf{l}_i'\lambda) \\ &\quad \times \mathbf{g}(\mathbf{z}_i'\alpha_1 + \delta_1 l_{i1}, \ldots, \mathbf{z}_i'\alpha_J + \delta_J l_{iJ})\big]\mathbf{h}(\mathbf{l}_i)d\mathbf{l}_i \\ &\approx \frac{1}{S}\sum_{s=1}^{S}\big[f(\mathbf{x}_i'\beta + \mathbf{d}_i'\gamma + \tilde{\mathbf{l}}_{is}'\lambda) \\ &\quad \times \mathbf{g}(\mathbf{z}_i'\alpha_1 + \delta_1\tilde{l}_{i1s}, \ldots, \mathbf{z}_i'\alpha_J + \delta_J\tilde{l}_{iJs})\big], \end{aligned} \tag{4.22}$$

where $\tilde{\mathbf{l}}_{is}$ is the sth draw (from a total of S draws) of a pseudo-random number from the density $\mathbf{h}$. Maximizing simulated log-likelihood is equivalent to maximizing the log-likelihood for S sufficiently large.

$$\begin{aligned}\ln l(y_i, \mathbf{d}_i|\mathbf{x}_i, \mathbf{z}_i) \approx \sum_{i=1}^{N}\ln\Bigg(\frac{1}{S}\sum_{s=1}^{S}\big[f(\mathbf{x}_i'\beta + \mathbf{d}_i'\gamma + \tilde{\mathbf{l}}_{is}'\lambda) \\ \times \mathbf{g}(\mathbf{z}_i'\alpha_1 + \delta_1\tilde{l}_{i1s}, \ldots, \mathbf{z}_i'\alpha_J + \delta_J\tilde{l}_{iJs})\big]\Bigg). \end{aligned} \tag{4.23}$$

For identification the scale of each choice equation should be normalized, and the covariances between choice equation errors be fixed. A natural set of normalization restrictions given by $\delta_{jk} = 0\ \forall j \neq k$, i.e., each choice is affected by a unique latent factor, and $\delta_{jj} = 1\,\forall j$, which normalizes the scale of each choice equation. This leads to an element in the covariance matrix being restricted to zero; see Deb and Trivedi (2006a) for details.

Under the unrealistic assumption of correct specification of the model, this approach will generate consistent, asymptotically normal, and efficient estimates. But the restrictions on preferences implied by the MMNL of choice are quite strong and not necessarily appropriate for all data sets. Estimation requires computer intensive simulation based methods that are discussed in Section 4.6.

4.4.4 Endogeneity in Two-Part Models

In considering endogeneity and self-selection in two-part models, we gain clarity by distinguishing carefully between several variants current in the literature. The baseline TPM model is that stated in Section 4.2; the first part is a model of dichotomous outcome whether the count is zero or positive, and the second part is a truncated count model, often the Poisson or NB, for positive counts. In this benchmark model the two parts are independent and all regressors are assumed to be strictly exogenous.

We now consider some extensions of the baseline. The first variant that we consider, referred to as TPM-S, arises when the independence assumption for

the two parts is dropped. Instead assume that there is a bivariate distribution of random variables (ν_1 ν_2), representing correlated unobserved factors that affect both the probability of the dichotomous outcome and the conditional count outcome. The two-parts are connected via unobserved heterogeneity. The resulting model is the count data analog of the classic Gronau-Heckman selection model applied to female labor force participation. It is also a special case of the model given in the previous section and can be formally derived by specializing Equations 4.18 to 4.20 to the case of one dichotomous variable and one truncated count distribution. Notice that in this variant the dichotomous endogenous variable will not appear as a regressor in the outcome equation. In practical application of the TPM-S model one is required to choose an appropriate distribution of unobserved heterogeneity. Greene (2007b) gives specific examples and relevant algebraic details. Following Terza (1998) he also provides the count data analog of Heckman two-step estimator.

A second variant of the two-part model is an extension of the TPM-S model described above as it also allows for dependence between the two parts of TPM and further allows for the presence of endogenous regressors in both parts. Hence we call this the TPM-ES model. If dependence between endogenous regressors and the outcome variable is introduced thorough latent factors as in Subsection 4.4.3, then such a model can be regarded a hybrid based on TPM-ES model and the latent factor model. Identification of such a model will require restrictions on the joint covariance matrix of errors, while simulation-based estimation appears to be a promising alternative.

The third and last variant of the TPM is a special case. It is obtained under the assumption that conditional on the inclusion of common endogenous regressor(s) in the two parts, plus the exogenous variables, the two parts are independent. We call this specification the TPM-E model. This assumption is not easy to justify, especially if endogeneity is introduced via dependent latent factors. However, if this assumption is accepted, estimation using moment-based IV estimation of each equation is feasible. Estimation of a class of binary outcome models with endogenous regressors is well established in the literature and has been incorporated in several software packages such as Stata. Both two-step sequential and ML estimators have been developed for the case of a continuous endogenous regressor; see Newey (1987). The estimator also assumes multivariate normality and homoscedasticity, and hence cannot be used for the case of an endogenous discrete regressor. Within the GMM framework the second part of the model will be based on the truncated moment condition

$$\mathrm{E}[y_i \exp(-\mathbf{x}_i'\beta) - 1 | \mathbf{z}_i, y_i > 0] = \mathbf{0}. \tag{4.24}$$

The restriction $y_i > 0$ is rarely exploited either in choosing the instruments or in estimation. Hence most of the discussion given in Subsection 4.4.1 remains relevant.

4.4.5 Bayesian Approaches to Endogeneity and Self-Selection

Modern Bayesian inference is attractive whenever the models are parametric and important features of models involve latent variables that can be simulated. There are two recent Bayesian analyses of endogeneity in count models that illustrate key features of such analyses; see Munkin and Trivedi (2003) and Deb, Munkin, and Trivedi (2006a). We sketch the structure of the model developed in the latter.

Deb, Munkin, and Trivedi (2006a) develop a Bayesian treatment of a more general potential outcome model to handle endogeneity of treatment in a count-data framework. For greater generality the entire outcome response function is allowed to differ between the treated and the nontreated groups. This extends the more usual selection model in which the treatment effect only enters through the intercept, as in Munkin and Trivedi (2003). This more general formulation uses the potential outcome model in which causal inference about the impact of treatment is based on a comparison of observed outcomes with constructed counterfactual outcomes. The specific variant of the potential outcome model used is often referred to as the "Roy model," which has been applied in many previous empirical studies of distribution of earnings, occupational choice, and so forth. The study extends the framework of the "Roy model" to nonnegative and integer-valued outcome variables and applies Bayesian estimation to obtain the full posterior distribution of a variety of treatment effects.

Define latent variable Z to measure the difference between the utility generated by two choices that reflect the benefits and the costs associated with them. Assume that Z is linear in the set of explanatory variables $\mathbf{W}$

$$Z = \mathbf{W}\alpha + u, \tag{4.25}$$

such that $d = 1$ if and only if $Z \geq \mathbf{0}$, and $d = 0$ if and only if $Z < \mathbf{0}$.

Assume that individuals choose between two regimes in which two different levels of utility are generated. As before latent variable Z, defined by Equation 4.25 where $u \sim \mathcal{N}(0, 1)$, measures the difference between the utility. In Munkin and Trivedi (2003) $d = 1$ means having private insurance (the treated state) and $d = 0$ means not having it (the untreated state). Two potential utilization variables Y_1, Y_2 are distributed as Poisson with means $\exp(\mu_1)$, $\exp(\mu_2)$, respectively. Variables μ_1, μ_2 are linear in the set of explanatory variables $\mathbf{X}$ and u such as

$$\mu_1 = \mathbf{X}\beta_1 + u\pi_1 + \varepsilon_1, \tag{4.26}$$

$$\mu_2 = \mathbf{X}\beta_2 + u\pi_2 + \varepsilon_2, \tag{4.27}$$

where $\mathrm{Cov}(u, \varepsilon_1|\mathbf{X}) = 0$, $\mathrm{Cov}(u, \varepsilon_2|\mathbf{X}) = 0$, and $\varepsilon = (\varepsilon_1, \varepsilon_2) \sim \mathcal{N}(\mathbf{0}, \Sigma)$, $\Sigma = \mathrm{diag}(\sigma_1, \sigma_2)$. The observability condition for Y is $Y = Y_1$ if $d = 1$ and $Y = Y_2$ if $d = 0$. The counted variable Y, representing utilization of medical services, is Poisson distributed with two different conditional means depending on the insurance status. Thus, there are two regimes generating count variables Y_1,

Y_2, but only one value is observed. Observe the restriction $\sigma_{12} = 0|\mathbf{X}, u$. This is imposed since the covariance parameter is unidentified in this model.

The standard Tanner–Wong data augmentation approach can be adapted to include latent variables μ_{1i}, μ_{2i}, Z_i in the parameter set making it a part of the posterior. Then the Bayesian MCMC approach can be used to obtain the posterior distribution of all parameters. A test to check the null hypothesis of no endogeneity is also feasible. Denote by M_1 the specification of the model that leaves parameters π_1 and π_2 unconstrained, and by M_0 the model that puts $\pi_1 = \pi_2 = 0$ constraint. Then a test of no endogeneity can be implemented using the Bayes factor $B_{0,1} = m(\mathbf{y}|M_0)/m(\mathbf{y}|M_1)$, where $m(\mathbf{y}|M)$ is the marginal likelihood of the model specification M.

In the case when the proportions of zero observations are so large that even extensions of the Poisson model that allow for overdispersion, such as negative binomial and the Poisson-lognormal models, do not provide an adequate fit, the ordered probit (OP) modeling approach might be an option. Munkin and Trivedi (2008) extend the OP model to allow for endogeneity of a set of categorical dummy covariates (e.g., types of health insurance plans), defined by a multinomial probit model (MNP). Let $\mathbf{d}_i = (d_{1i}, d_{2i}, \ldots, d_{J-1i})$ be binary random variables for individual i ($i = 1, \ldots, N$) choosing category j ($j = 1, \ldots, J$) (category J is the baseline) such that $d_{ji} = 1$ if alternative j is chosen and $d_{ji} = 0$ otherwise. The MNP model is defined using the multinomial latent variable structure which represents gains in utility received from the choices, relative to the utility received from choosing alternative J. Let the $(J-1) \times 1$ random vector $\mathbf{Z}_i$ be defined as

$$\mathbf{Z}_i = \mathbf{W}_i\alpha + \varepsilon_i,$$

where $\mathbf{W}_i$ is a matrix of exogenous regressors, such that

$$d_{ji} = \prod_{l=1}^{J} I_{[0,+\infty)}\left(Z_{ji} - Z_{li}\right), \quad j = 1, \ldots, J,$$

where $Z_{Ji} = 0$ and $I_{[0,+\infty)}$ is the indicator function for the set $[0, +\infty)$. The distribution of the error term ε_i is $(J-1)$-variate normal $\mathcal{N}(\mathbf{0}, \Sigma)$. For identification it is customary to restrict the leading diagonal element of Σ to unity.

To model the ordered dependent variable it is assumed that there is another latent variable Y_i^* that depends on the outcomes of $\mathbf{d}_i$ such that

$$Y_i^* = \mathbf{X}_i\beta + \mathbf{d}_i\rho + u_i,$$

where $\mathbf{X}_i$ is a vector of exogenous regressors, and ρ is a $(J-1) \times 1$ parameter vector. Define Y_i as

$$Y_i = \sum_{m=1}^{M} m I_{[\tau_{m-1}, \tau_m)}\left(Y_i^*\right),$$

where $\tau_0, \tau_1, \ldots, \tau_M$ are threshold parameters and $m = 1, \ldots, M$. For identification, it is standard to set $\tau_0 = -\infty$ and $\tau_M = \infty$ and additionally restrict

$\tau_1 = 0$. The choice of insurance is potentially endogenous to utilization and this endogeneity is modeled through correlation between u_i and ε_i, assuming that they are jointly normally distributed with variance of u_i restricted for identification since Y_i^* is latent; see Deb, Munkin, and Trivedi (2006b).

Munkin and Trivedi (2009) extend the Ordered Probit model with Endogenous Selection to allow for a covariate such as income to enter the insurance equation nonparametrically. The insurance equation is specified as

$$Z_i = f(s_i) + \mathbf{W}_i\alpha + \varepsilon_i, \tag{4.28}$$

where $\mathbf{W}_i$ is a vector of regressors, α is a conformable vector of parameters, and the distribution of the error term ε_i is $\mathcal{N}(0, 1)$. Function $f(.)$ is unknown and s_i is income of individual i. The data are sorted by values of s so that s_1 is the lowest level of income and s_N is the largest. The main assumption made on function $f(s_i)$ is that it is smooth such that it is differentiable and its slope changes slowly with s_i such that, for a given constant C, $|f(s_i) - f(s_{i-1})| \leq C|s_i - s_{i-1}|$ — a condition which covers a wide range of functions.

Economic theory predicts that risk-averse individuals prefer to purchase insurance against catastrophic or simply costly evens because they value eliminating risk more than money at sufficiently high wealth levels. This is modeled by assuming that a risk-averse individual's utility is a monotonically increasing function of wealth with diminishing marginal returns. This is certainly true for general medical insurance when liabilities could easily exceed any reasonable levels. However, in the context of dental insurance the potential losses have reasonable bounds. Munkin and Trivedi (2009) find strong evidence of diminishing marginal returns of income on dental insurance status and even a nonmonotonic pattern.

4.5 Panel Data

We begin with a model for scalar dependent variable y_{it} with regressors $\mathbf{x}_{it}$, where i denotes the individual and t denotes time. We will restrict our coverage to the case of t small, usually referred to as "short panel," which is also of most interest in microeconometrics. Assuming multiplicative individual scale effects applied to exponential function

$$\mathrm{E}[y_{it}|\alpha_i, \mathbf{x}_{it}] = \alpha_i \exp(\mathbf{x}_{it}'\beta), \tag{4.29}$$

As $\mathbf{x}_{it}$ includes an intercept, α_i may be interpreted as a deviation from 1 because $\mathrm{E}(\alpha_i|x) = 1$.

In the standard case in econometrics the time interval is fixed and the data are equi-spaced through time. However, the panel framework can also cover the case where the data are simply repeated events and not necessarily equispaced through time. An example of such data is the number of epileptic

seizures during a two-week period preceding each of four consecutive clinical visits; see Diggle et al. (2002).

4.5.1 Pooled or Population-Averaged (PA) Models

Pooling occurs when the observations $y_{it}|\alpha_i, \mathbf{x}_{it}$ are treated as independent, after assuming $\alpha_i = \alpha$. Consequently cross-section observations can be "stacked" and cross-section estimation methods can then be applied.

The assumption that data are poolable is strong. For parametric models it is assumed that the marginal density for a single (i, t) pair,

$$f(y_{it}|\mathbf{x}_{it}) = f(\alpha + \mathbf{x}_{it}'\beta, \gamma), \tag{4.30}$$

is correctly specified, regardless of the (unspecified) form of the joint density

$$f(y_{it}, \ldots, y_{iT}|\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}, \beta, \gamma).$$

The pooled model, also called the population-averaged (PA) model, is easily estimated. A panel-robust or cluster-robust (with clustering on i) estimator of the covariance matrix can then be applied to correct standard errors for any dependence over time for given individual. This approach is the analog of pooled OLS for linear models.

The pooled model for the exponential conditional mean specifies $\mathrm{E}[y_{it}|\mathbf{x}_{it}] = \exp(\alpha + \mathbf{x}_{it}'\beta)$. Potential efficiency gains can be realized by taking into account dependence over time. In the statistics literature such an estimator is constructed for the class of generalized linear models (GLM) that includes the Poisson regression. Essentially this requires that estimation be based on weighted first-order moment conditions to account for correlation over t, given i, while consistency is ensured provided the conditional mean is correctly specified as $\mathrm{E}[y_{it}|\mathbf{x}_{it}] = \exp(\alpha + \mathbf{x}_{it}'\beta) \equiv g(\mathbf{x}_{it}, \beta)$. The efficient GMM estimator, known in the statistics literature as the population-averaged model, or generalized estimating equations (GEE) estimator (see Diggle et al. [2002]), is based on the conditional moment restrictions, stacked over all T observations,

$$\mathrm{E}[\mathbf{y}_i - \mathbf{g}_i(\beta)|\mathbf{X}_i] = \mathbf{0}, \tag{4.31}$$

where $\mathbf{g}_i(\beta) = [g(\mathbf{x}_{i1}, \beta), \ldots, g(\mathbf{x}_{iT}, \beta)]'$ and $\mathbf{X}_i = [\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}]'$. The optimally weighted unconditional moment condition is

$$\mathrm{E}\left[\frac{\partial \mathbf{g}_i'(\beta)}{\partial \beta}\{\mathrm{V}[\mathbf{y}_i|\mathbf{X}_i]\}^{-1}(\mathbf{y}_i - \mathbf{g}_i(\beta))\right] = \mathbf{0}. \tag{4.32}$$

Given Σ_i a working variance matrix for $\mathrm{V}[\mathbf{y}_i|\mathbf{X}_i]$, the moment condition becomes

$$\sum_{i=1}^{N} \frac{\partial \mathbf{g}_i'(\beta)}{\partial \beta}\Sigma_i^{-1}(\mathbf{y}_i - \mathbf{g}_i(\beta)) = \mathbf{0}. \tag{4.33}$$

The asymptotic variance matrix, which can be derived using standard GEE/GMM theory (see CT, 2005, Chapter 23.2), is robust to misspecification of Σ_i. For the case of strictly exogenous regressors the GEE methodology is not strictly speaking "recent," although it is more readily implementable nowadays because of software developments.

While the foregoing analysis applies to the case of additive errors, there are multiplicative versions of moment conditions (as detailed in Subsection 4.4.1) that will lead to different estimators. Finally, in the case of endogenous regressors, the choice of the optimal GMM estimator is more complicated as it depends upon the choice of optimal instruments; if $\mathbf{z}_i$ defines a vector of valid instruments, then so does any function $h(\mathbf{z}_i)$.

Given its strong restrictions, the GEE approach connects straightforwardly with the GMM/IV approach used for handling endogenous regressors. To cover the case of endogenous regressors we simply rewrite the previous moment condition as $\mathrm{E}[\mathbf{y}_i - \mathbf{g}_i(\beta)|\mathbf{Z}_i] = \mathbf{0}$, where $\mathbf{Z}_i = [\mathbf{z}_{i1}, \ldots, \mathbf{z}_{iT}]'$ are appropriate instruments.

Because of the greater potential for having omitted factors in panel models of observational data, fixed and random effect panel count models have relatively greater credibility than the above PA model. The strong restrictions of the pooled panel model are relaxed in different ways by random and fixed effects models. The recent developments have impacted the random effects panel models more than the fixed effect models, in part because computational advances have made them more accessible.

4.5.2 Random-Effects Models

A random-effects (RE) model treats the individual-specific effect α_i as an unobserved random variable with specified mixing distribution $g(\alpha_i|\gamma)$, similar to that considered for cross-section models of Section 4.2. Then α_i is eliminated by integrating over this distribution. Specifically the unconditional density for the ith observation is

$$f(y_{it}, \ldots, y_{iT_i}|\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT_i}, \beta, \gamma, \eta) = \int \left[\prod_{t=1}^{T_i} f(y_{it}|\mathbf{x}_{it}, \alpha_i, \beta, \gamma) \right] g(\alpha_i|\eta) d\alpha_i. \tag{4.34}$$

For some combinations of $\{f(\cdot), g(\cdot)\}$ this integral usually has analytical solution. However, if randomness is restricted to the intercept only, then numerical integration is also feasible as only univariate integration is required. The RE approach, when extended to both intercept and slope parameters, becomes computationally more demanding.

As in the cross-section case, the negative binomial panel model can be derived under two assumptions: first, y_{ij} has Poisson distribution conditional on μ_i, and second, μ_i are i.i.d. gamma distributed with mean μ and variance $\alpha\mu^2$. Then, unconditionally $y_{ij} \sim NB(\mu_i, \mu_i + \alpha\mu_i^2)$. Although this model is easy to estimate using standard software packages, it has the obvious limitation

that it requires a strong distributional assumption for the random intercept and it is only useful if the regressors in the mean function $\mu_i = \exp(\mathbf{x}_i'\beta)$ do not vary over time. The second assumption is frequently violated.

Morton (1987) relaxed both assumptions of the preceding paragraph and proposed a GEE-type estimator for the following exponential mean with multiplicative heterogeneity model: $\mathrm{E}[y_{it}|\mathbf{x}_{it}, \nu_i] = \exp(\mathbf{x}_{it}'\beta)\nu_i$; $\mathrm{Var}[y_{it}|\nu_i] = \phi\mathrm{E}[y_{it}|\mathbf{x}_{it}, \nu_i]$; $\mathrm{E}[\nu_i] = 1$ and $\mathrm{Var}[\nu_i] = \alpha$. These assumptions imply $\mathrm{E}[y_{it}|\mathbf{x}_{it}] = \exp(\mathbf{x}_{it}'\beta)$ and $\mathrm{Var}[y_{it}] = \phi\mu_{it} + \alpha\mu_{it}^2$. A GEE-type estimator based on Equation 4.33 is straight-forward to construct; see Diggle et al. (2002).

Another example is Breslow and Clayton (1993) who consider the specification

$$\ln\{\mathrm{E}[y_{it}|\mathbf{x}_{it}, z_{it}]\} = \mathbf{x}_{it}'\beta + \gamma_{1t} + \gamma_{2t}z_{it},$$

where the intercept and slope coefficients $(\gamma_{1t}, \gamma_{2t})$ are assumed to be bivariate normal distributed. Whereas regular numerical integration estimation for this can be unstable, adaptive quadrature methods have been found to be more robust; see Rabe-Hesketh, Skrondal, and Pickles (2002).

A number of authors have suggested a further extension of the RE models mentioned above; see Chib, Greenberg, and Winkelmann (1998). The assumptions of this model are: 1. $y_{it}|\mathbf{x}_{it}, \mathbf{b}_i \sim \mathcal{P}(\mu_{it})$; $\mu_{it} = \mathrm{E}[y_{it}|\mathbf{x}_{it}'\beta + \mathbf{w}_{it}'\mathbf{b}_i]$; and $\mathbf{b}_i \sim \mathcal{N}[\mathbf{b}^*, \Sigma_\mathbf{b}]$ where $(\mathbf{x}_{it}')$ and $(\mathbf{w}_{it}')$ are vectors of regressors with no common elements and only the latter have random coefficients. This model has an interesting feature that the contribution of random effect is not constant for a given i. However, it is fully parametric and maximum likelihood is computationally demanding. Chib, Greenberg, and Winkelmann (1998) use Markov chain Monte Carlo to obtain the posterior distribution of the parameters.

A potential limitation of the foregoing RE panel models is that they may not generate sufficient flexibility in the specification of the conditional mean function. Such flexibility can be obtained using a finite mixture or latent class specification of random effects and the mixing can be with respect to the intercept only, or all the parameters of the model. Specifically, consider the model

$$f(y_{it}|\beta, \pi) = \sum_{j=1}^{m} \pi_j(z_{it}|\gamma) f_j(y_{it}|\mathbf{x}_{it}, \beta_j), \quad 0 < \pi_j(\cdot) < 1, \sum_{j=1}^{m} \pi_j(\cdot) = 1 \tag{4.35}$$

where for generality the mixing probabilities are parameterized as functions of observable variables z_{it} and parameters γ, and the j-component conditional densities may be any convenient parametric distributions, e.g., the Poisson or negative binomial, each with its own conditional mean function and (if relevant) a variance parameter. In this case individual effects are approximated using a distribution with finite number of discrete mass points that can be interpreted as the number of "types." Such a specification offers considerable flexibility, albeit at the cost of potential over-parametrization. Such a model is a straightforward extension of the finite mixture cross-section model. Bago d'Uva (2005) uses the finite mixture of the pooled negative binomial in her

study of primary care using the British Household Panel Survey; Bago d'Uva (2006) exploits the panel structure of the Rand Health Insurance Experiment data to estimate a latent class hurdle panel model of doctor visits.

The RE model has different conditional mean from that for pooled and population-averaged models, unless the random individual effects are additive or multiplicative. So, unlike the linear case, pooled estimation in nonlinear models leads to inconsistent parameter estimates if instead the assumed random-effects model is appropriate, and vice-versa.

4.5.3 Fixed-Effects Models

Given the conditional mean specification

$$\mathrm{E}[y_{it}|\alpha_i, \mathbf{x}_{it}] = \alpha_i \exp(\mathbf{x}_{it}'\beta) = \alpha_i \mu_{it}, \tag{4.36}$$

a fixed-effects (FE) model treats α_i as an unobserved random variable that may be correlated with the regressors $\mathbf{x}_{it}$. It is known that maximum likelihood or moment-based estimation of both the population-averaged Poisson model and the RE Poisson model will not identify the β if the FE specification is correct. Econometricians often favor the fixed effects specification over the RE model. If the FE model is appropriate then a fixed-effects estimator should be used, but it may not be available if the problem of incidental parameters cannot be solved. Therefore, we examine this issue in the following section.

4.5.3.1 Maximum Likelihood Estimation

Whether, given short panels, joint estimation of the fixed effects $\alpha = (\alpha_1, \ldots, \alpha_N)$ and β is feasible is the first important issue. Under the assumption of strict exogeneity of $\mathbf{x}_{it}$, the basic result that there is no incidental parameter problem for the Poisson panel regression is now established and well understood (CT 1998; Lancaster 2000; Windmeijer 2008). Consequently, corresponding to the fixed effects, one can introduce N dummy variables in the Poisson conditional mean function and estimate (α, β) by maximum likelihood. This will increase the dimensionality of the estimation problem. Alternatively, the conditional likelihood principle may be used to eliminate α and to condense the log-likelihood in terms of β only. However, maximizing the condensed likelihood will yield estimates identical to those from the full likelihood. Table 4.2 displays the first order condition for FE Poisson MLE of β, which can be compared with the pooled Poisson first-order condition to see how the fixed effects change the estimator. The difference is that μ_{it} in the pooled model is replaced by $\mu_{it}\bar{y}_i/\bar{\mu}_i$ in the FE Poisson MLE. The multiplicative factor $\bar{y}_i/\bar{\mu}_i$ is simply the ML estimator of α_i; this means the first-order condition is based on the likelihood concentrated with respect to α_i.

The result about the incidental parameter problem for the Poisson FE model does not extend to the fixed effects NB2 model (whose variance function is quadratic in the conditional mean) if the fixed effects parameters enter multiplicatively through the conditional mean specification. This fact is confusing

TABLE 4.2

Selected Moment Conditions for Panel Count Models

Model	Moment or Model Specification	Estimating Equations or Moment Condition
Pooled Poisson	$\mathrm{E}[y_{it}\|\mathbf{x}_{it}] = \exp(\mathbf{x}_{it}'\beta)$,	$\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbf{x}_{it}(y_{it}-\mu_{it})=\mathbf{0}$ where $\mu_{it}=\exp(\mathbf{x}_{it}'\beta)$
Pop. averaged		$\rho_{ts}=\mathrm{Cor}[(y_{it}-\exp(\mathbf{x}_{it}'\beta))$ $(y_{is}-\exp(\mathbf{x}_{is}'\beta))]$.
Poisson RE	$\mathrm{E}[y_{it}\|\alpha_i,\mathbf{x}_{it}] = \alpha_i\exp(\mathbf{x}_{it}'\beta)$,	$\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbf{x}_{it}\left(y_{it}-\mu_{it}\frac{\bar{y}_i+\eta/T}{\bar{\mu}_i+\eta/T}\right)=\mathbf{0}$ $\bar{\mu}_i = T^{-1}\sum_t\exp(\mathbf{x}_{it}'\beta);\ \eta=\mathrm{Var}(\alpha_i)$
Poisson FE	$\mathrm{E}[y_{it}\|\alpha_i,\mathbf{x}_{it}] = \alpha_i\exp(\mathbf{x}_{it}'\beta)$	$\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbf{x}_{it}\left(y_{it}-\mu_{it}\frac{\bar{y}_i}{\bar{\mu}_i}\right)=\mathbf{0}$,
GMM (Windmeijer)	$y_{it}=\exp(\mathbf{x}_{it}'\beta+\alpha_i)u_{it}$,	$\sum_{i=1}^{N}\sum_{t=1}^{T}\left[y_{it}\frac{\mu_{it-1}}{\mu_{it}}-y_{it-1}\|\mathbf{x}_i^{t-1}\right]=\mathbf{0}$
Strict exog	$\mathrm{E}[\mathbf{x}_{it}u_{it+j}]=0,\ j\geq 0$	
Predetermined reg.	$\mathrm{E}[\mathbf{x}_{it}u_{it-s}]\neq 0,\ s\geq 1$	
GMM (Wooldridge)	$\mathrm{E}\left[\frac{y_{it}}{\mu_{it}}-\frac{y_{it-1}}{\mu_{it-1}}\|\mathbf{x}_i^{t-1})\right]=0$	$\sum_{i=1}^{N}\sum_{t=1}^{T}\left[\frac{y_{it}}{\mu_{it}}-\frac{y_{it-1}}{\mu_{it-1}}\|\mathbf{x}_i^{t-1})\right]=\mathbf{0}$
GMM (Chamberlain)	$\mathrm{E}\left[y_{it}\frac{\mu_{it-1}}{\mu_{it}}-y_{it-1}\|\mathbf{x}_i^{t-1})\right]=0$	$\sum_{i=1}^{N}\sum_{t=1}^{T}\left[y_{it}\frac{\mu_{it-1}}{\mu_{it}}-y_{it-1}\|\mathbf{x}_i^{t-1})\right]=\mathbf{0}$
GMM/endog	$\mathrm{E}\left[\frac{y_{it}}{\mu_{it}}-\frac{y_{it-1}}{\mu_{it-1}}\|\mathbf{x}_i^{t-2})\right]=0$	$\sum_{i=1}^{N}\sum_{t=1}^{T}\left[y_{it}\frac{\mu_{it-1}}{\mu_{it}}-y_{it-1}\|\mathbf{x}_i^{t-2}\right]=\mathbf{0}$
Dynamic feedabck	$y_{it}=\theta y_{it-1}+\exp(\mathbf{x}_{it}'\beta+\alpha_i)$ $+u_{it}$,	$\mathrm{E}\Big[(y_{it}-\theta y_{it-1})\frac{\mu_{it-1}}{\mu_{it}}-$ $-(y_{it-1}-\theta y_{it-2})\|y_{it-2},\mathbf{x}_i^{t-1}\Big]=\mathbf{0}$

for many practitioners who observe the availability of the fixed effects NB option in several commercial computer packages. Greene (2007b) provides a good exposition of this issue. He points out that the option in the packages is that of Hausman, Hall, and Griliches (1984) who specified a variant of the "fixed effects negative binomial" (FENB) distribution in which the variance function is linear in the conditional mean; that is, $\mathrm{Var}[y_{it}|\mathbf{x}_{it}] = (1+\alpha_i)\mathrm{E}[y_{it}|\mathbf{x}_{it}]$, so the variance is a scale factor multiplied by the conditional mean, and the fixed effects parameters enter the model through the scaling factor. This is the NB model with linear variance (or NB1), not that with a quadratic variance (or NB2 formulation). As fixed effects come through the variance function, not the conditional mean, this is clearly a different formulation from the Poisson fixed effects model. Given that the two formulations are not nested, it is not clear how one should compare FE Poisson and this particular variant of the FENB. Greene (2007b) discusses related issues in the context of an empirical example.

4.5.3.2 Moment Function Estimation

Modern literature considers and sometimes favors the use of moment-based estimators that may be potentially more robust than the MLE. The starting point here is a moment condition model. Following Chamberlain (1992), and mimicing the differencing transformations used to eliminate nuisance parameters in linear models, there has been an attempt to obtain moment condition models based on quasi-differencing transformations that eliminate fixed effects; see Wooldridge (1999, 2002). This step is then followed by application of one of the several available variants of the GMM estimation, such as two-step GMM or continuously updated GMM. Windmeier (2008) provides a good survey of the approach for the Poisson panel model.

Windmeier (2008) considers the following alternative formulations:

$$y_{it} = \exp(\mathbf{x}'_{it}\beta + \alpha_i)u_{it}, \tag{4.37}$$

$$y_{it} = \exp(\mathbf{x}'_{it}\beta + \alpha_i) + u_{it}, \tag{4.38}$$

where, in the first case $E(u_{it}) = 1$, the $\mathbf{x}_{it}$ are predetermined with respect to u_{it}, and u_{it} are serially uncorrelated and independent of α_i. The table lists the implied restriction. A quasi-differencing transformation eliminates the fixed effects and generates moment conditions whose form depend on whether we start with Equation 4.37 or 4.38. Several variants are shown in Table 4.2 and they can be used in GMM estimation. Of course, these moment conditions only provide a starting point and important issues remain about the performance of alternative variants or the best variants to use. Windmeier (2008) discusses the issues and provides a Monte Carlo evaluation.

It is conceivable that a fixed effects–type formulation may adequately account for overdispersion of counts. But there are other complications that generate overdispersion in other ways, e.g., excess zeros and fat tails. At present little is known about the performance of moment-based estimators when the d.g.p. deviates significantly from the Poisson-type behavior. Moment-based models do not exploit the integer-valued aspect of the dependent variable. Whether this results in significant efficiency loss — and if so, when — is a topic that deserves future investigation.

4.5.4 Conditionally Correlated Random Effects

The standard random effect panel model assumes that α_i and $\mathbf{x}_{it}$ are uncorrelated. Instead we can relax this and assume that they are conditionally correlated. This idea, originally developed in the context of a linear panel model by Mundlak (1978) and Chamberlain (1982), can be interpreted as intermediate between fixed and random effects. That is, if the correlation between α_i and the regressors can be controlled by adding some suitable "sufficient" statistic for the regressors, then the remaining unobserved heterogeneity can be treated as random and uncorrelated with the regressors. While in principle we may introduce a subset of regressors, in practice it is more parsimonious to introduce time-averaged values of time-varying regressors. This is

the conditionally correlated random (CCR) effects model. This formulation allows for correlation by assuming a relationship of the form

$$\alpha_i = \bar{\mathbf{x}}_i'\lambda + \varepsilon_i, \tag{4.39}$$

where $\bar{\mathbf{x}}$ denotes the time-average of the time-varying exogenous variables and ε_i may be interpreted as unobserved heterogeneity uncorrelated with the regressors. Substituting this into the above formulation essentially introduces no additional problems except that the averages change when new data are added. To use the standard RE framework, however, we need to make an assumption about the distribution of ε_t and this will usually lead to an integral that would need evaluating. Estimation and inference in the pooled Poisson or NLS model can proceed as before. This formulation can also be used when dynamics are present in the model.

Because the CCR formulation is intermediate between the FE and RE models, it may serve as a useful substitute for not being able to deal with FE in some specifications. For example, a panel version of the hurdle model with FE is rarely used as the fixed effects cannot be easily eliminated. In such a case the CCR specification is feasible.

4.5.5 Dynamic Panels

As in the case of linear models, inclusion of lagged values is appropriate in some empirical models. An example is the use of past research and development expenditure when modeling the number of patents, see Hausman, Hall, and Griliches (1984). When lagged exogenous variables are used, no new modeling issues arise from their presence. However, to model lagged dependence more flexibly and more parsimoniously, the use of lagged dependent variables y_{t-j} ($j \geq 1$) as regressors is attractive, but it introduces additional complications that have been studied in the literature on autoregressive models of counts (see CT [1998], Chapters 7.4 and 7.5). Introducing autoregressive dependence through the exponential mean specification leads to a specification of the type

$$\mathrm{E}[y_{it}|\mathbf{x}_{it}, y_{it-1}, \alpha_i] = \exp(\gamma y_{it-1} + \mathbf{x}_{it}'\beta + \alpha_i), \tag{4.40}$$

where α_i is the individual-specific effect. If the α_i are uncorrelated with the regressors, and further if parametric assumptions are to be avoided, then this model can be estimated using either the nonlinear least squares or pooled Poisson MLE. In either case it is desirable to use the robust variance formula.

The estimation of a dynamic panel model requires additional assumptions about the relationship between the initial observations ("initial conditions") y_0 and the α_i. For example, using the CCR model we could write $\alpha_i = y_0'\delta + \bar{\mathbf{x}}_i'\lambda + \varepsilon_i$ where y_0 is an initial condition. Then maximum likelihood estimation could proceed by treating the initial condition as given. The alternative of taking the initial condition as random, specifying a distribution for it, and then integrating out the condition is an approach that has

been suggested for other dynamic panel models, and it is computationally more demanding; see Stewart (2007). Under the assumption that the initial conditions are nonrandom, the standard random effects conditional maximum likelihood approach identifies the parameters of interest. For a class of nonlinear dynamic panel models, including the Poisson model, Wooldridge (2005) analyzes this model which conditions the joint distribution on the initial conditions.

The inclusion of lagged y_{it} inside the exponential mean function introduces potentially sharp discontinuities that may result in a poor fit to the data. It is not the case that this will always happen, but it might when the range of counts is very wide. Crepon and Duguet (1997) proposed using a better starting point in a dynamic fixed effects panel model; they specified the model as

$$y_{it} = h(y_{it-1}, \theta) \exp(\mathbf{x}_{it}'\beta + \alpha_i) + u_{it} \tag{4.41}$$

where the function $h(y_{it-1}, \theta)$ parametrizes the dependence on lagged values of y_{it}. Crepon and Duguet (1997) suggested switching functions to allow lagged zero values to have a different effect from positive values. Blundell, Griffith, and Windmeijer (2002) proposed a linear feedback model with multiplicative fixed effect α_i,

$$y_{it} = \theta y_{it-1} + \exp(\mathbf{x}_{it}'\beta + \alpha_i) + u_{it}, \tag{4.42}$$

but where the lagged value enters linearly. This formulation avoids awkward discontinuities and is related to the integer valued autoregressive (INAR) models. A quasi-differencing transformation can be applied to generate a suitable estimating equation. Table 4.2 shows the estimating equation obtained using a Chamberlain-type quasi-differencing transformation. Consistent GMM estimation here depends upon the assumption that regressors are predetermined. Combining this with the CCR assumption about α_i is straight forward.

Currently the published literature does not provide detailed information on the performance of the available estimators for dynamic panels. Their development is in early stages and, not surprisingly, we are unaware of commercial software to handle such models.

4.6 Multivariate Models

Multivariate count regression models, especially its bivariate variant, are of empirical interest in many contexts. In the simplest case one may be interested in the dependence structure between counts $y_1, \ldots, y_m$, conditional on vectors of exogenous variables $\mathbf{x}_1, \ldots, \mathbf{x}_m$, $m \geq 2$. For example, y_1 denotes the number of prescribed and y_2 the number of nonprescribed medications taken by individuals over a fixed period.

4.6.1 Moment-Based Models

The simplest and attractive semiparametric approach here follows Delgado (1992); it simply extends the seemingly unrelated regressions (SUR) for linear models to the case of multivariate exponential regression. For example, in the bivariate case we specify $\mathrm{E}[\mathbf{y}_1|\mathbf{x}_1] = \exp(\mathbf{x}_1'\beta_1)$ and $\mathrm{E}[\mathbf{y}_2|\mathbf{x}_2] = \exp(\mathbf{x}_2'\beta_2)$, assume additive errors and then apply nonlinear least squares, but estimate variances using the heteroscedasticity-robust variance estimator supported by many software packages. This is simply nonlinear SUR and is easily extended to several equations. It is an attractive approach when all conditional means have exponential mean specifications and the joint distribution is not desired. It also permits a very flexible covariance structure and its asymptotic theory is well established. Tests of cross-equation restrictions are easy to implement.

An extension of the model would include a specification for variances and covariance. For example, we could specify $\mathrm{V}[\mathbf{y}_j|\mathbf{x}_j] = \alpha_j \exp(\mathbf{x}_j'\beta_j)$, $j = 1, 2$, and $\mathrm{Cov}[\mathbf{y}_1, \mathbf{y}_2|\mathbf{x}_1, \mathbf{x}_1] = \rho \times \exp(\mathbf{x}_1'\beta_1)^{1/2} \exp(\mathbf{x}_1'\beta_2)^{1/2}$. This specification is similar to univariate Poisson quasi-likelihood except improved efficiency is possible using a generalized estimating equations estimator.

4.6.2 Likelihood-Based Models

At issue is the joint distribution of $(y_1, y_2|\mathbf{x}_1, \mathbf{x}_2)$. A different data situation is one in which y_1 and y_2 are paired observations that are jointly distributed, whose marginal distributions $f_1(y_1|\mathbf{x}_1)$ and $f_2(y_2|\mathbf{x}_2)$ are parametrically specified, but our interest is in some function of y_1 and y_2. They could be data on twins, spouses, or paired organs (kidneys, lungs, eyes), and the interest lies in studying and modeling the difference. When the bivariate distribution of (y_1, y_2) is known, standard methods can be used to derive the distribution of any continuous function of the variables, say $H(y_1, y_2)$.

A problem arises, however, when an analytical expression for the joint distribution is either not available at all or is available in an explicit form only under some restrictive assumptions. This situation arises in case of multivariate Poisson and negative binomial distributions that are only appropriate for positive dependence between counts, thus lacking generality. Unrestricted multivariate distributions of discrete outcomes often do not have closed form expressions, see Marshall and Olkin (1990), CT (1998), and Munkin and Trivedi (1999). The first issue to consider is how to generate flexible specifications of multivariate count models. The second issue concerns estimation and inference.

4.6.2.1 Latent Factor Models

One fruitful way to generate flexible dependence structures between counts is to begin by specifying latent factor models. Munkin and Trivedi (1999) generate a more flexible dependence structure using a correlated unobserved

heterogeneity model. Suppose y_1 and y_2 are, respectively, $\mathcal{P}(\mu_1|\nu_1)$ and $\mathcal{P}(\mu_2|\nu_2)$

$$\mathrm{E}[y_j|\mathbf{x}_j,\nu_j] = \mu_j = \exp(\beta_{0j} + \lambda_j\nu_j + \mathbf{x}_j'\beta_j), \quad j = 1, 2 \tag{4.43}$$

where ν_1 and ν_2 represent correlated latent factors or unobserved heterogeneity and (λ_1, λ_2) are factor loadings. Dependence is induced if ν_1 and ν_2 are correlated. Assume (ν_1, ν_2) to be bivariate normal distributed with correlation ρ, $0 \leq \rho \leq 1$. Integrating out (ν_1, ν_2), we obtain the joint distribution

$$f(y_1, y_2|\mathbf{x}_1, \mathbf{x}_2, \nu_1, \nu_2) = \int f_1(\mathbf{y}_1|\mathbf{x}_1, \nu_1) f_2(\mathbf{y}_2|\mathbf{x}_2, \nu_2) g(\nu_1, \nu_2) d\nu_1 d\nu_2, \tag{4.44}$$

where the right-hand side can be replaced by simulation-based numerical approximation

$$\frac{1}{S}\sum_{s=1}^{S} f_1\big(\mathbf{y}_1|\mathbf{x}_1, \nu_1^{(s)}\big) f_2\big(\mathbf{y}_2|\mathbf{x}_2, \nu_2^{(s)}\big), \tag{4.45}$$

The method of simulation-based maximum likelihood (SMLE) estimates the unknown parameters using the likelihood based on such an approximation. As shown in Munkin and Trivedi (1999), while SMLE of $(\beta_{01}, \beta_1, \beta_{02}, \beta_2, \lambda_1, \lambda_2)$ is feasible it is not computationally straightforward. Recently two alternatives to SMLE have emerged. The first uses Bayesian Monte Carlo Markov Chain (MCMC) approach to estimation; see Chib and Winkelmann (2001). MCMC estimation is illustrated in Subsection 4.6.3. The second uses copulas to generate a joint distribution whose parameters can be estimated without simulation.

4.6.2.2 Copulas

Copula-based joint estimation is based on Sklar's theorem which provides a method of generating joint distributions by combining marginal distributions using a copula. Given a continuous m-variate distribution function $F(y_1, \ldots, y_m)$ with univariate marginal distributions $F_1(y_1), \ldots, F_m(y_m)$ and inverse (quantile) functions $F_1^{-1}, \ldots, F_m^{-1}$, then $y_1 = F_1^{-1}(u_1) \sim F_1, \ldots, y_m = F_m^{-1}(u_m) \sim F_m$, where $u_1, \ldots, u_m$ are uniformly distributed variates. By Sklar's theorem, an m-copula is an m-dimensional distribution function with all m univariate margins being $U(0, 1)$, i.e.,

$$F(y_1, \ldots, y_m) = F(F_1^{-1}(u_1), \ldots, F_m^{-1}(u_m)) = C(u_1, \ldots, u_m; \theta), \tag{4.46}$$

is the unique copula associated with the distribution function. Here $C(\cdot)$ is a given functional form of a joint distribution function and θ is a dependence parameter. Zero dependence implies that the joint distribution is the product of marginals. A leading example is a Gaussian copula based on any relevant marginal such as the Poisson.

Sklar's theorem implies that copulas provide a "recipe" to derive joint distributions when only marginal distributions are given. The approach is attractive because copulas (1) provide a fairly general approach to joint modeling of count data; (2) neatly separate the inference about marginal distribution from inference on dependence; (3) represent a method for deriving joint distributions given the fixed marginals such as Poisson and negative binomial; (4) in a bivariate case copulas can be used to define nonparametric measures of dependence that can capture asymmetric (tail) dependence as well as correlation or linear association; (4) are easier to estimate than multivariate latent factor models with unobserved heterogeneity. However, copulas and latent factor models are closely related; see Trivedi and Zimmer (2007) and Zimmer and Trivedi (2006).

The steps involved in copula modeling is specification of marginal distributions and a copula. There are many possible choices of copula functional forms, see Nelsen (2006). The resulting model can be estimated by a variety of methods such as joint maximum likelihood of all parameters, or two-step estimation in which marginal models are estimated first and θ is estimated at the second step. For details see Trivedi and Zimmer (2007).

An example of copula estimation is Cameron et al. (2004) who use the copula framework to analyze the empirical distribution of two counted measures, y_1 denoting self-reported doctor visits, and y_2 denoting independent report of doctor visits. They derive the distribution of $y_1 - y_2$, by first obtaining the joint distribution $f[y_1, y_2]$. Zimmer and Trivedi (2006) use a trivariate copula framework to develop a joint distribution of two counted outcomes and one binary treatment variable.

There is growing interest in Bayesian analysis of copulas. A recent example is Pitt, Chan, and Kohn (2006), who use a Gaussian copula to model the joint distribution of six count measures of health care. Using a multivariate density of the Gaussian copula Pitt, Chan, and Kohn develop a MCMC algorithm for estimating the posterior distribution for discrete marginals, which is then applied to the case where marginal densities are zero-inflated geometric distributions.

4.7 Simulation-Based Estimation

Simulation-based estimation methods, both classical and Bayesian, deal with distributions that do not have closed form solutions. Such distributions are usually generated when general assumptions are made on unobservable variables that need to be integrated out. The classical estimation methods include both parametric and semiparametric approaches. Hinde (1982) and Gouriéroux and Monfort (1991) discuss a parametric Simulated Maximum Likelihood (SML) approach to estimation of mixed-Poisson regression models. Application to some random effects panel count models has been

implemented by Crepon and Duguet (1997). Delgado (1992) treats a multivariate count model as a multivariate nonlinear model and suggests a semiparametric generalized least squares estimator. Gurmu and Elder (2007) develop a flexible semiparametric specification using generalized Laguerre polynomials, and propose a semiparametric estimation method without distributional specification of the unobservable heterogeneity. Another approach (Cameron and Johansson 1997) is based on series expansion methods putting forward a squared polynomial series expansion.

Bayesian estimation of both univariate and multivariate Poisson models is a straightforward Gibbs sampler in the case when regressors do not enter the mean parameters of the Poisson distribution. However, since an objective of economists is to calculate various marginal and treatment effects, such covariates must be introduced. This leads to a necessity to use Metropolis-Hastings steps in the MCMC algorithms. In the era when high speed computers were not available Bayesian estimation of various models relied on deriving a closed form posterior distributions whenever possible. When such closed forms do not exist as in the case of the Poisson model, the posterior can be numerically approximated (El-Sayyad 1973). However, since an inexpensive computer power became available a path of utilizing MCMC methods has been taken. Chib, Greenberg, and Winkelmann (1998) propose algorithms based on MCMC methods to deal with panel count data models with random effects. Chib and Winkelmann (2001) develop an MCMC algorithm of a multivariate correlated count data model. Munkin and Trivedi (2003) extend a count data model to account for a binary endogenous treatment variable. Deb, Munkin, and Trivedi (2006a) introduce a Roy-type count model with the proposed algorithm being more efficient (with respect to computational time and convergence) than the existing MCMC algorithms dealing with Poisson-lognormal densities.

4.7.1 The Poisson-Lognormal Model

Whereas the Poisson-gamma mixture model, i.e., the NB distribution, has proved very popular in application, different distributional assumptions on unobserved heterogeneity might be more consistent with real data. One such example is the Poisson lognormal model.

The Poisson lognormal model is a continuous mixture in which the marginal count distribution is still assumed to be Poisson and the distribution of the multiplicative unobserved heterogeneity term ν is lognormal. Let us reparameterize ν such that $\nu = \exp(\varepsilon)$, where $\varepsilon \sim \mathcal{N}\left(\mathbf{0}, \sigma^2\right)$, and let the mean of the marginal Poisson distribution be a function of a vector of exogenous variables $\mathbf{X}$.

The count variables y is distributed as Poisson with mean $\exp(\mu)$, where μ is linear in $\mathbf{X}$ and ε

$$\mu = \mathbf{X}\beta + \varepsilon, \tag{4.47}$$

where $\text{Cov}(\varepsilon|\mathbf{X}) = 0$. Then conditionally on unobserved heterogeneity term ε, the marginal count distribution is defined as

$$f(y|\mathbf{X}, \beta, \varepsilon) = \frac{\exp\left[-\exp(\mathbf{X}\beta + \varepsilon)\right]\exp\left[y(\mathbf{X}\beta + \varepsilon)\right]}{y!}. \tag{4.48}$$

The unconditional density $f(y|\mathbf{X}, \beta, \sigma)$ does not have a closed form since the integral

$$\int_{-\infty}^{\infty} f(y|\mathbf{X}, \beta, \varepsilon) f(\varepsilon|\sigma)\, d\varepsilon \tag{4.49}$$

cannot be solved. Since in many applications the lognormal distribution is a more appealing assumption on unobserved heterogeneity than gamma, a reliable estimation method of such a model is needed. Estimation by Gaussian quadrature is very feasible; Winkelmann (2004) provides a good illustration.

4.7.2 SML Estimation

Assume that we have N independent observations. An SML estimator of $\theta = (\beta, \sigma)$ is defined as

$$\widehat{\theta}_{SN} = \arg\max_{\theta} \sum_{i=1}^{N} \log\left\{\frac{1}{S}\sum_{s=1}^{S} f(y_i|\mathbf{X}_i, \beta, \varepsilon_i^s)\right\}, \tag{4.50}$$

where ε_i^s $(s = 1, \ldots, S)$ are drawn from density $f(\varepsilon|\sigma)$. In our case this density depends on unknown parameter σ. Instead of introducing an importance sampling function we reparameterize the model such that $\mu = \mathbf{X}\beta + \sigma u$, where $u \sim \mathcal{N}(\mathbf{0}, 1)$. Then

$$\begin{aligned} f(y|\mathbf{X}, \beta, \sigma) = \int_{-\infty}^{\infty} & \frac{\exp([-\exp(\mathbf{X}\beta + \sigma u)]\exp[y(\mathbf{X}\beta + \sigma u)]}{y!} \\ & \times \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) du. \end{aligned} \tag{4.51}$$

In this example the standard normal density of u is a natural candidate for the importance sampling function. Then the SML estimates maximize

$$\sum_{i=1}^{N} \log\left\{\frac{1}{S}\sum_{s=1}^{S} \frac{\exp(\left[-\exp\left(\mathbf{X}_i\beta + \sigma u_i^s\right)\right]\exp\left[y_i\left(\mathbf{X}_i\beta + \sigma u_i^s\right)\right]}{y!}\right\}, \quad j = 1, 2 \tag{4.52}$$

where u_i^s are drawn from $N[0, 1]$.

Since log is an increasing function, the sum over i and log do not commute. Then if S is fixed and N tends to infinity $\widehat{\theta}_{SN}$ is not consistent. If both S and N tend to infinity then the SML estimator is consistent.

4.7.3 MCMC Estimation

Next we discuss the choice of the priors and outline the MCMC algorithm. For each observation i derive the joint density of the observable data and latent variables. We adopt the Tanner–Wong data augmentation approach and include latent variables μ_i $(i = 1, \ldots, N)$ in the parameter set making it a part of the posterior. Conditional on μ_i the full conditional density of β is a tractable normal distribution.

Denote $\Delta_i = (\mathbf{X}_i, \beta, \sigma)$. Then the joint density of the observable data and latent variables for observation i is

$$f(y_i, \mu_i|\Delta_i) = \frac{\exp[y_i\mu_i - \exp(\mu_i)]}{y_i!} \frac{1}{\sqrt{2\pi\sigma^2}} \exp[-0.5\sigma^{-2}(\mu_i - \mathbf{X}_i\beta)^2]. \quad (4.53)$$

The posterior density kernel is the product of $f(y_i, \mu_i|\Delta_i)$ for all N observations and the prior densities of the parameters.

We choose a normal prior for parameter β, center it at zero and choose relatively large variance

$$\beta \sim \mathcal{N}(\mathbf{0}_k, 10\mathbf{I}_k). \quad (4.54)$$

The priors for the variance parameter is

$$\sigma^{-2} \sim \mathcal{G}\left(\frac{n}{2}, \left(\frac{c}{2}\right)^{-1}\right) \text{ where } n = 5 \text{ and } c = 10.$$

First, we block the parameters as $\mu_i, \beta, \sigma^{-2}$. The steps of the MCMC algorithm are the following:

1. The full conditional density for μ_i is proportional to

$$p(\mu_i|\Delta_i) = \frac{\exp[y_i\mu_i - \exp(\mu_i)]}{y_i!} \exp[-0.5\sigma^{-2}(\mu_i - \mathbf{X}_i\beta)^2]. \quad (4.55)$$

Sample μ_i using the Metropolis–Hasting algorithm with normal distribution centered at the modal value of the full conditional density for the proposal density. Let

$$\widehat{\mu}_i = \arg\max \log p(\mu_i|\Delta_i) \quad (4.56)$$

and $\mathbf{V}_{\widehat{\mu}_i} = -(\mathbf{H}_{\widehat{\mu}_i})^{-1}$ be the negative inverse of the Hessian of $\log p(\mu_i|\Delta_i)$ evaluated at the mode $\widehat{\mu}_i$. Choose the proposal distribution $q(\mu_i) = \phi(\mu_i|\widehat{\mu}_i, \mathbf{V}_{\widehat{\mu}_i})$. When a proposal value μ_i^* is drawn, the chain moves to the proposal value with probability

$$\alpha(\mu_i, \mu_i^*) = \min\left\{\frac{p(\mu_i^*|\Delta_i)q(\mu_i)}{p(\mu_i|\Delta_i)q(\mu_i^*)}, 1\right\}. \quad (4.57)$$

If the proposal value is rejected, the next state of the chain is at the current value μ_i.

2. Specify prior distributions $\beta \sim \mathcal{N}\left[\underline{\beta}, \underline{H}_{\beta}^{-1}\right]$. The the conditional distribution of β is $\beta \sim \mathcal{N}[\overline{\beta}, \overline{\mathbf{H}}_{\beta}^{-1}]$ where

$$\overline{\mathbf{H}}_{\beta} = \underline{\mathbf{H}}_{\beta} + \sum_{i=1}^{N} \mathbf{X}_i' \sigma^{-2} \mathbf{X}_i \tag{4.58}$$

$$\overline{\beta} = \overline{\mathbf{H}}_{\beta}^{-1} \left[\underline{\mathbf{H}}_{\beta} \underline{\beta} + \sum_{i=1}^{N} \mathbf{X}_i' \sigma^{-2} \mu_i \right]. \tag{4.59}$$

3. Finally, specify the prior $\sigma^{-2} \sim \mathcal{G}(n/2, (c/2)^{-1})$. Then the full conditional of σ^{-2} is

$$\mathcal{G}\left(\frac{n+N}{2}, \left[\frac{c}{2} + \sum_{i=1}^{N} \frac{(\mu_i - \mathbf{X}_i \beta)^2}{2} \right]^{-1} \right). \tag{4.60}$$

This concludes the MCMC algorithm.

4.7.4 A Numerical Example

To examine properties of our SML estimator and MCMC algorithm and their performance, we generate several artificial data sets. In this section we report our experience based on one specific data generating process (d.g.p.). We generate 1000 observations using the following structure with assigned parameter values: $X_i = (1, x_i)$ and $x_i \sim \mathcal{N}(0, 1)$; $\beta = (2, 1)$, $\sigma = 1$. Such parameter values generate a count variable with mean of 19. The priors for parameters are selected to be uninformative but still proper, i.e., $\beta \sim \mathcal{N}(\mathbf{0}, 10 I_2)$ and $\sigma^{-1} \sim \mathcal{G}\left(\frac{n}{2}, \left(\frac{c}{2}\right)^{-1}\right)$ with $n = 5$ and $c = 10$.

Table 4.3 gives SML estimates and the posterior means and standard deviations for the parameters based on 10,000 replications preceded by 1000 replications of the burn-in phase. It also gives the true values of the parameters in the d.g.p. As can be seen from the table the true values of the parameters fall close to the centers of the estimated confidence intervals. However, if the true values of β is selected such that the mean of the count variable is increased to 50, the estimates of the SML estimator display a considerable bias when the number of simulations is limited to $S = 500$.

TABLE 4.3

MCMC Estimation for Generated Data

Parameter	True Value of d.g.p.	MCMC	SML
β_0 (*Constant*)	2	1.984	1.970
		0.038	0.036
β_1 (x)	1	0.990	0.915
		0.039	0.027
σ	1	1.128	1.019
		0.064	0.026

4.7.5 Simulation-Based Estimation of Latent Factor Model

We now consider some issues in the estimation of the latent factor model of Subsection 4.6.1. The literature indicates that S should increase faster than $\sqrt{N}$, but this does not give explicit guidance in choosing S. In practice some tests of convergence should be applied to ensure that S was set sufficiently high. Using a small number of draws (often 50–100) works well for models such as the mixed multinomial logit, multinomial probit, etc. However, more draws are required for models with endogenous regressors. Thus computation can be quite burdensome if the standard methods are used. For the model described in Subsection 4.4.3, Deb and Trivedi (2006b) find the standard simulation methods to be quite slow. They adapt a simulation acceleration technique that uses quasi-random draws based on Halton sequences (Bhat 2001; Train 2002). This method, instead of using S pseudo-random points, makes draws based on a nonrandom selection of points within the domain of integration. Under suitable regularity conditions, the integration error using pseudo-random sequences is in the order of N^{-1} as compared to pseudo-random sequences where the convergence rate is $N^{-1/2}$ (Bhat 2001). For variance estimation, they use the robust Huber–White formula.

4.8 Software Matters

In the past decade the scope of applying count data models has been greatly enhanced by availability of good software and fast computers. Leading microeconometric software packages such as Limdep, SAS, Stata, and TSP provide a good coverage of the basic count model estimation for single equation and Poisson-type panel data models. See Greene (2007a) for details of Limdep, and Stata documentation for coverage of Stata's official commands; also see Kitazawa (2000) and Romeu (2004). The present authors are especially familiar with Stata official estimation commands. The Poisson, ZIP, NB, and ZINB are covered in the Stata reference manuals. Stata commands support calculation of marginal effects for most models. Researchers should also be aware that there are other add-on Stata commands that can be downloaded from Statistical Software Components Internet site at Boston College Department of Economics. These include commands for estimating hurdle and finite mixture models due to Deb (2007), goodness-of-fit and model evaluation commands due to Long and Freese (2006), quantile count regression commands due to Miranda (2006), and commands due to Deb and Trivedi (2006b) for simulation-based estimation of multinomial latent factor model discussed in Subsection 4.4.3. Stata 11, released in late 2009, facilitates implementing GMM estimation of cross-section and panel data models based on the exponential mean specification.

4.8.1 Issues with Bayesian Estimation

The main computational difficulty with the simulated maximum likelihood approach is the fact that when the number of simulations is small the parameters estimates are biased. This is true for even simple one equation models. When the model becomes multivariate and multidimensional a much larger number of simulations is required for consistent estimation. Sometimes it can be very time consuming with the computational time increasing exponentially with the number of parameters. In Bayesian Markov chain Monte Carlo the computational time increases proportionally to the dimension of the model. Besides, the approach does not suffer from the bias problem of the SML. However, there are computational problems with the Markov chain Monte Carlo methods as well. Such problems arise when the produced Markov chains display a high level of serial correlation leading to the posterior distribution being saturated in a closed neighborhood with the Markov chain not visiting the entire support of the posterior distribution. When the serial correlation is high but reasonably smaller, the solution is to use a relatively larger number of replications for a precise estimation of the posterior. However, when the serial correlations are close to one such a problem must have a model specific solution.

Bayesian model specification requires a choice of priors which can result in a completely different posterior. When improper priors are selected this can lead to improper posterior. In general, Bayesian modeling does not restrict itself to only customized models and new programs must be written for various model specifications. Many programs for the well-developed existing models are written in MATLAB. Koop, Porier, and Tobias (2007) give an excellent overview of different methods and models and provide a rich library of programs. This book can serve as a good MATLAB reference for researchers dealing with Bayesian modeling and estimation.

References

Bago d'Uva, T. 2005. Latent class models for use of primary care: evidence from a British panel. *Health Economics* 14: 873–892.

Bago d'Uva, T. 2006. Latent class models for use of health care. *Health Economics* 15: 329–343.

Besag, J. 1974. Spatial interaction and the statistical analysis of lattice systems. *Journal of Royal Statistical Society 36B*: 192–225.

Bhat, C. R. 2001. Quasi-random maximum simulated likelihood estimation of the mixed multinomial logit model. *Transportation Research: Part B* 35: 677–693.

Blundell, R., R. Griffith, and F. Windmeijer. 2002. Individual effects and dynamics in count data models. *Journal of Econometrics* 102: 113–131.

Bohning, D., and R. Kuhnert. 2006. Equivalence of truncated count mixture distributions and mixtures of truncated count distributions, *Biometrics* 62(4): 1207–1215.

Breslow, N. E., and D. G. Clayton. 1993. Approximate inference in generalized linear mixed models. *Journal of American Statistical Association* 88: 9–25.

Cameron, A. C., and P. Johansson. 1997. Count data regressions using series expansions with applications. *Journal of Applied Econometrics* 12(3): 203–223.

Cameron, A. C., T. Li, P. K. Trivedi, and D. M. Zimmer. 2004. Modeling the differences in counted outcomes using bivariate copula models: with application to mismeasured counts. *Econometrics Journal* 7(2): 566–584.

Cameron, A. C., and P. K. Trivedi. 1998. *Regression Analysis of Count Data*. New York: Cambridge University Press.

Cameron, A. C., and P. K. Trivedi. 2005. *Microeconometrics: Methods and Applications*. Cambridge, U.K.: Cambridge University Press.

Cameron, A. C., and P. K. Trivedi. 2009. *Microeconometrics Using Stata*. College Station, TX: Stata Press.

Chamberlain, G. 1982. Multivariate regression models for panel data. *Journal of Econometrics* 18: 5–46.

Chamberlain, G. 1992. Comment: sequential moment restrictions in panel data. *Journal of Business and Economic Statistics* 10: 20–26.

Chang, F. R., and P. K. Trivedi. 2003. Economics of self-medication: theory and evidence. *Health Economics* 12: 721–739.

Chib, S., E. Greenberg, and R. Winkelmann. 1998. Posterior simulation and Bayes factor 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.

Crepon, B., and E. Duguet. 1997. Research and development, competition and innovation: pseudo-maximum likelihood and simulated maximum likelihood method applied to count data models with heterogeneity. *Journal of Econometrics* 79: 355–378.

Davidson, R., and J. G. MacKinnon. 2004. *Econometric Theory and Methods*, Oxford, U.K.: Oxford University Press.

Davis, R. A., W. T. M. Dunsmuir, and S. B. Streett. 2003. Observation-driven models for Poisson counts. *Biometrika* 90: 777–790.

Deb, P. 2007. FMM: Stata module to estimate finite mixture models. Statistical Software Components S456895. Boston College Department of Economics.

Deb, P., M. K. Munkin, and P. K. Trivedi. 2006a. Private insurance, selection, and the health care use: a Bayesian analysis of a Roy-type Model. *Journal of Business and Economic Statistics* 24: 403–415.

Deb, P., M. K. Munkin, and P. K. Trivedi. 2006b. Bayesian analysis of the two-part model with endogeneity: application to health care expenditure. *Journal of Applied Econometrics* 21(6): 1081–1099.

Deb, P., and P. K. Trivedi. 1997. Demand for medical care by the elderly: a finite mixture approach. *Journal of Applied Econometrics* 12: 313–326.

Deb, P., and P. K. Trivedi. 2002. The structure of demand for medical care: latent class versus two-part models. *Journal of Health Economics* 21: 601–625.

Deb, P., and P. K. Trivedi. 2006a. Specification and simulated likelihood estimation of a non-normal treatment-outcome model with selection: application to health care utilization. *Econometrics Journal* 9: 307–331.

Deb, P., and P. K. Trivedi. 2006b. Maximum simulated likelihood estimation of a negative-binomial regression model with multinomial endogenous treatment. *Stata Journal* 6: 1–10.

Delgado, M. A. 1992. Semiparametric generalized least squares in the multivariate nonlinear regression model. *Econometric Theory* 8: 203–222.

Demidenko, E. 2007. Poisson regression for clustered data. *International Statistical Review* 75(1): 96–113.

Diggle, P., P. Heagerty, K. Y. Liang, and S. Zeger. 2002. *Analysis of Longitudinal Data.* Oxford, U.K.: Oxford University Press.

El-Sayyad, G. M. 1973. Bayesian and classical analysis of poisson regression. *Journal of the Royal Statistical Society*, Series B (Methodological) 35(3): 445–451.

Fruhwirth-Schnatter, S. 2006. *Finite Mixture and Markov Switching Models*. New York: Springer-Verlag.

Gouriéroux, C., and A. Monfort. 1991. Simulation based inference in models with heterogeneity. *Annales d'Economie et de Statistique* 20/21: 69–107.

Gourieroux, C., and A. Monfort. 1997. *Simulation Based Econometric Methods*. Oxford, U.K.: Oxford University Press.

Greene, W. H. 2007a. *LIMDEP 9.0 Reference Guide*. Plainview, NY: Econometric Software, Inc.

Greene, W. H. 2007b. Functional form and heterogeneity in models for count data. *Foundations and Trends in Econometrics* 1(2): 113–218.

Griffith, D. A., and R. Haining. 2006. Beyond mule kicks: the Poisson distribution in geographical analysis. *Geographical Analysis* 38: 123–139.

Guo, J. Q., and P. K. Trivedi. 2002. Flexible parametric distributions for long-tailed patent count distributions. *Oxford Bulletin of Economics and Statistics* 64: 63–82.

Gurmu, S., and J. Elder. 2007. A simple bivariate count data regression model. *Economics Bulletin* 3 (11): 1–10.

Gurmu, S., and P. K. Trivedi. 1996. Excess zeros in count models for recreational trips. *Journal of Business and Economic Statistics* 14: 469–477.

Hardin, J. W., H. Schmiediche, and R. A. Carroll. 2003. Instrumental variables, bootstrapping, and generalized linear models. *Stata Journal* 3: 351–360.

Hausman, J. A., B. H. Hall, and Z. Griliches. 1984. Econometric models for count data with an application to the patents–R and D relationship. *Econometrica* 52: 909–938.

Hinde, J. 1982. Compound Poisson regression models. In R. Gilchrist ed., 109-121, GLIM 82: *Proceedings of the International Conference on Generalized Linear Models*. New York: Springer-Verlag.

Jung, R. C., M. Kukuk, and R. Liesenfeld. 2006. Time series of count data: modeling, estimation and diagnostics. *Computational Statistics & Data Analysis* 51: 2350–2364.

Kaiser, M., and N. Cressie. 1997. Modeling Poisson variables with positive spatial dependence. *Statistics and Probability Letters* 35: 423–32.

Karlis, D., and E. Xekalaki. 1998. Minimum Hellinger distance estimation for Poisson mixtures. *Computational Statistics and Data Analysis* 29: 81–103.

Kitazawa, Y. 2000. TSP procedures for count panel data estimation. Fukuoka, Japan: Kyushu Sangyo University.

Koenker, R. 2005. *Quantile Regression*. New York: Cambridge University Press.

Koop, G., D. J. Poirier, and J. L. Tobias. 2007. *Bayesian Econometric Methods*. Volume 7 of Econometric Exercises Series. New York: Cambridge University Press.

Lancaster, T. 2000. The incidental parameters problem since 1948. *Journal of Econometrics* 95: 391–414.

Long, J. S., and J. Freese. 2006. *Regression Models for Categorical Dependent Variables Using Stata*, 2nd ed. College Station, TX: Stata Press.

Lourenco, O. D., and P. L. Ferreira. 2005. Utilization of public health centres in Portugal: effect of time costs and other determinants. Finite mixture models applied to truncated samples. *Health Economics* 14: 939–953.

Lu, Z., Y. V. Hui, and A. H. Lee. 2003. Minimum Hellinger distance estimation for finite mixtures of Poisson regression models and its applications. *Biometrics* 59(4): 1016–1026.

MacDonald, I. L., and W. Zucchini. 1997. *Hidden Markov and Other Models for Discrete-Valued Time Series*. London: Chapman & Hall.

Machado, J., and J. Santos Silva. 2005. Quantiles for counts. *Journal of American Statistical Association* 100: 1226–1237.

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, *Topics in Statistical Dependence*, IMS Lecture Notes-Monograph Series, Volume 16: 371–393.

Miranda, A. 2006. QCOUNT: Stata program to fit quantile regression models for count data. Statistical Software Components S456714. Boston College Department of Economics.

Miranda, A. 2008. Planned fertility and family background: a quantile regression for counts analysis. *Journal of Population Economics* 21: 67–81.

Morton, R. 1987. A generalized linear model with nested strata of extra-Poisson variation. *Biometrika* 74: 247–257.

Mullahy, J. 1997. Instrumental variable estimation of Poisson regression models: application to models of cigarette smoking behavior. *Review of Economics and Statistics* 79: 586–593.

Mundlak, Y. 1978. On the pooling of time series and cross section data. *Econometrica* 56: 69–86.

Munkin, M., and P. K. Trivedi. 1999. Simulated maximum likelihood estimation of multivariate mixed-Poisson regression models, with application. *Econometric Journal* 1: 1–21.

Munkin, M. K., and P. K. Trivedi. 2003. Bayesian analysis of self-selection model with multiple outcomes using simulation-based estimation: an application to the demand for healthcare. *Journal of Econometrics* 114: 197–220.

Munkin, M. K., and P. K. Trivedi, 2008. Bayesian analysis of the ordered Probit model with endogenous selection, *Journal of Econometrics* 143: 334–348.

Munkin, M. K., and P. K. Trivedi, 2009. A Bayesian analysis of the OPES Model with a non-parametric component: application to dental insurance and dental care. Forthcoming in *Advances in Econometrics, Volume 23: Bayesian Econometrics*, edited by Siddhartha Chib, Gary Koop, and Bill Griffiths. Elsevier Press.

Nelsen, R. B. 2006. *An Introduction to Copulas*. 2nd ed. New York: Springer.

Newey, W. 1987. Efficient estimation of limited dependent variable models with endogenous explanatory variables. *Journal of Econometrics* 36: 231–250.

Pitt, M., D. Chan, and R. Kohn. 2006. Efficient Bayesian inference for Gaussian copula regression. *Biometrika* 93: 537–554.

Rabe-Hesketh, S., A. Skrondal, and A. Pickles. 2002. Reliable estimation of generalized linear mixed models using adaptive quadrature. *Stata Journal* 2: 1–21.

Romeu, A. 2004. *ExpEnd:* Gauss code for panel count data models. *Journal of Applied Econometrics* 19: 429–434.

Skrondal, A., and S. Rabe-Hesketh. 2004. *Generalized Latent Variable Modeling: Multilevel, Longitudinal and Structural Equation Models*. London: Chapman & Hall.

Stewart, M. 2007. The inter-related dynamics of unemployment and low-wage employment. *Journal of Applied Econometrics* 22(3): 511–531.

Terza, J. 1998. Estimating count data models with endogenous switching: sample selection and endogenous switching effects. *Journal of Econometrics* 84: 129–139.

Train, K. 2002. *Discrete Choice Methods with Simulation.* New York: Cambridge University Press.

Trivedi, P. K., and D. M. Zimmer. 2007. Copula modeling: an introduction for practitioners. *Foundations and Trends in Econometrics* 1(1): 1–110.

Vuong, Q. 1989. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica* 57: 307–333.

Wang, K., K. K. W. Yau, and A. H. Lee. 2002. A hierarchical Poisson mixture regression model to analyze maternity length of hospital stay. *Statistics in Medicine* 21: 3639–3654.

Windmeijer, F. 2008. GMM for panel count data models. *Advanced Studies in Theoretical and Applied Econometrics* 46: 603–624.

Windmeijer, F., and J. M. C. Santos Silva. 1997. Endogeneity in count data models. *Journal of Applied Econometrics* 12: 281–294.

Winkelmann, R. 2004. Health care reform and the number of doctor visits – an econometric analysis. *Journal of Applied Econometrics* 19: 455–472.

Winkelmann, R. 2005. *Econometric Analysis of Count Data*. 5th ed. Berlin: Springer-Verlag.

Winkelmann, R. 2006. Reforming health care: Evidence from quantile regressions for counts. *Journal of Health Economics* 25: 131–145.

Wooldridge, J. M. 1997. Multiplicative panel data models without the strict exogeneity assumption. *Econometric Theory* 13: 667–678.

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*, 2001. Cambridge, MA: MIT Press.

Wooldridge, J. M. 2005. Simple solutions to the initial conditions problem in dynamic, nonlinear panel data models with unobserved heterogeneity. *Journal of Applied Econometrics* 20: 39–54.

Xiang, L., K. K. W. Yau, Y. Van Hui, and A. H. Lee. 2008. Minimum Hellinger distance estimation for k-component Poisson mixture with random effects. *Biometrics* 64(2): 508–518.

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(1): 63–76.