# Essentials of Count Data Regression

A. Colin Cameron
Email: accameron@ucdavis.edu

Pravin K. Trivedi
Email: trivedi@indiana.edu

June 30 1999

## 1. Introduction

In many economic contexts the dependent or response variable of interest ($y$) is a nonnegative integer or count which we wish to explain or analyze in terms of a set of covariates ($\mathbf{x}$). Unlike the classical regression model, the response variable is discrete with a distribution that places probability mass at nonnegative integer values only. Regression models for counts, like other limited or discrete dependent variable models such as the logit and probit, are nonlinear with many properties and special features intimately connected to discreteness and nonlinearity.

Let us consider some examples from microeconometrics, beginning with samples of independent cross-section observations. Fertility studies often model the number of live births over a specified age interval of the mother, with interest in analyzing its variation in terms of, say, mother's schooling, age, and household income (Winkelmann, 1995). Accident analysis studies model airline safety, for example, as measured by the number of accidents experienced by an airline over some period, and seek to determine its relationship to airline profitability and other measures of the financial health of the airline (Rose, 1990). Recreational demand studies seek to place a value on natural resources such as national forests by modeling the number of trips to a recreational site (Gurmu and Trivedi, 1996). Health demand studies model data on the number of times that individuals consume a health service, such as visits to a doctor or days in hospital in the past year (Cameron, Trivedi, Milne and Piggott, 1986), and estimate the impact of health status and health insurance.

Examples of count data regression based on time series and panel data are also available. A time series example is the annual number of bank failures over some period, which may be analyzed using explanatory variables such as bank profitability, corporate profitability, and bank borrowings from the Federal Reserve Bank (Davutyan, 1989). A panel data example that has attracted much attention in the industrial organization literature on the benefits of research and development expenditures is the number of patents received annually by firms (Hausman, Hall, and Griliches, 1984).

In some cases, such as number of births, the count is the variable of ultimate interest. In other cases, such as medical demand and results of research and development expenditure, the variable of ultimate interest is continuous, often expenditures or receipts measured in dollars, but the best data available are instead a count.

In all cases the data are concentrated on a few small discrete values, say 0, 1 and 2; skewed to the left; and intrinsically heteroskedastic with variance increasing with the mean.

In many examples, such as number of births, virtually all the data are restricted to single digits, and the mean number of events is quite low. But in other cases such as number of patents the tail can be very long with, say, one-quarter of the sample being awarded no patents while one firm is awarded 400 patents.

These features motivate the application of special methods and models for count regression. There are two ways to proceed. The first approach is a fully parametric one that completely specifies the distribution of the data, fully respecting the restriction of $y$ to nonnegative integer values. The second approach is a mean-variance approach, which specifies the conditional mean to be nonnegative, and specifies the conditional variance to be a function of the conditional mean.

These approaches are presented for cross-section data in Sections 2 to 4. Section 2 details the Poisson regression model. This model is often too restrictive and other, more commonly-used, fully parametric count models are presented in Section 3. Less-used alternative parametric approaches for counts, such as discrete choice models and duration models, are also presented in this section. The partially parametric approach of modeling the conditional mean and conditional variance is detailed in Section 4. Extensions to other types of data, notably time series, multivariate and panel data, are given in Section 5. In Section 6 practical recommendations are provided. For pedagogical reasons the Poisson regression model for cross-section data is presented in some detail. The other models, many superior to Poisson, are presented in less detail for space reasons. For more complete treatment see Cameron and Trivedi (1998) and the guide to further reading in Section 7.

## 2. Poisson Regression

The Poisson is the starting point for count data analysis, though it is often inadequate. In Sections 2.1-2.3 we present the Poisson regression model and estimation by maximum likelihood, interpretation of the estimated coefficients, and extensions to truncated and censored data. Limitations of the Poisson model, notably overdispersion, are presented in Section 2.4.

### 2.1. Poisson MLE

The natural stochastic model for counts is a Poisson point process for the occurrence of the event of interest. This implies a Poisson distribution for the number of occurrences of the event, with density

$$\Pr[Y = y] = \frac{e^{-\mu}\mu^y}{y!}, \qquad y = 0, 1, 2, ..., \tag{2.1}$$

where $\mu$ is the intensity or rate parameter. We refer to the distribution as $\mathsf{P}[\mu]$. The first two moments are

$$\begin{aligned} \mathsf{E}[Y] &= \mu, \\ \mathsf{V}[Y] &= \mu. \end{aligned} \tag{2.2}$$

This shows the well-known equality of mean and variance property of the Poisson distribution.

By introducing the observation subscript $i$, attached to both $y$ and $\mu$, the framework is extended to non-iid data. The *Poisson regression model* is derived from the Poisson distribution by parameterizing the relation between the mean parameter $\mu$ and covariates (regressors) $\mathbf{x}$. The standard assumption is to use the exponential mean parameterization,

$$\mu_i = \exp(\mathbf{x}_i'\boldsymbol{\beta}), \quad i = 1, ..., n, \tag{2.3}$$

where by assumption there are $k$ linearly independent covariates, usually including a constant. Because $\mathsf{V}[y_i|\mathbf{x}_i] = \exp(\mathbf{x}_i'\boldsymbol{\beta})$, by (2.2) and (2.3), the Poisson regression is intrinsically heteroskedastic.

Given (2.1) and (2.3) and the assumption that the observations $(y_i|\mathbf{x}_i)$ are independent, the most natural estimator is maximum likelihood (ML). The log-likelihood function is

$$\ln \mathsf{L}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \{y_i \mathbf{x}_i'\boldsymbol{\beta} - \exp(\mathbf{x}_i'\boldsymbol{\beta}) - \ln y_i!\}. \tag{2.4}$$

The Poisson MLE, denoted $\widehat{\boldsymbol{\beta}}_P$, is the solution to $k$ nonlinear equations corresponding to the first-order condition for maximum likelihood,

$$\sum_{i=1}^{n} (y_i - \exp(\mathbf{x}_i'\boldsymbol{\beta}))\mathbf{x}_i = \mathbf{0}. \tag{2.5}$$

If $\mathbf{x}_i$ includes a constant term then the residuals $y_i - \exp(\mathbf{x}_i'\boldsymbol{\beta})$ sum to zero by (2.5). The log-likelihood function is globally concave; hence solving these equations by Gauss-Newton or Newton-Raphson iterative algorithm yields unique parameters estimates.

By standard maximum likelihood theory of correctly specified models, the estimator $\widehat{\boldsymbol{\beta}}_P$ is consistent for $\boldsymbol{\beta}$ and asymptotically normal with the sample covariance matrix

$$\mathsf{V}[\widehat{\boldsymbol{\beta}}_P] = \left( \sum_{i=1}^{n} \mu_i \mathbf{x}_i \mathbf{x}_i' \right)^{-1}, \tag{2.6}$$

in the case where $\mu_i$ is of the exponential form (2.3). In practice an alternative more general form for the variance matrix should be used; see Section 4.1.

## 2.2. Interpretation of Regression Coefficients

For linear models, with $\mathsf{E}[y|\mathbf{x}] = \mathbf{x}'\boldsymbol{\beta}$, the coefficients $\boldsymbol{\beta}$ are readily interpreted as the effect of a one-unit change in regressors on the conditional mean. For nonlinear models this interpretation needs to be modified. For any model with exponential conditional mean, differentiation yields

$$\frac{\partial \mathsf{E}[y|\mathbf{x}]}{\partial x_j} = \beta_j \exp(\mathbf{x}'\boldsymbol{\beta}), \tag{2.7}$$

where the scalar $x_j$ denotes the $j^{th}$ regressor. For example, if $\widehat{\beta}_j = 0.25$ and $\exp(\mathbf{x}_i'\widehat{\boldsymbol{\beta}}) = 3$, then a one unit change in the $j^{th}$ regressor increases the expectation of $y$ by 0.75 units. This partial response depends upon $\exp(\mathbf{x}_i'\widehat{\boldsymbol{\beta}})$ which is expected to vary across individuals. It is easy to see that $\beta_j$ measures the relative change in $\mathsf{E}[y|\mathbf{x}]$ induced by a unit change in $x_j$. If $x_j$ is measured on log-scale, $\beta_j$ is an elasticity.

For purposes of reporting a single response value, a good candidate is an estimate of the *average response*, $\frac{1}{n}\sum_{i=1}^{n}\partial\mathsf{E}[y_i|\mathbf{x}_i]/\partial x_{ij} = \widehat{\beta}_j \times \frac{1}{n}\sum_{i=1}^{n}\exp(\mathbf{x}_i'\widehat{\boldsymbol{\beta}})$. For Poisson regression models with intercept included, this can be shown to simplify to $\widehat{\beta}_j\overline{y}$.

Another consequence of (2.7) is that if, say, $\beta_j$ is twice as large as $\beta_k$, then the effect of changing the $j^{th}$ regressor by one unit is twice that of changing the $k^{th}$ regressor by one unit.

## 2.3. Truncation and Censoring

In some studies, inclusion in the sample requires that sampled individuals have been engaged in the activity of interest. Then the count data are *truncated*, as the data are observed only over part of the range of the response variable. Examples of truncated counts include the number of bus trips made per week in surveys taken on buses, the number of shopping trips made by individuals sampled at a mall, and the number of unemployment spells among a pool of unemployed. In all these cases we do not observe zero counts, so the data are said to be zero-truncated, or more generally left-truncated. Right truncation results from loss of observations greater than some specified value.

Truncation leads to inconsistent parameter estimates unless the likelihood function is suitably modified. Consider the case of zero truncation. Let $f(y|\boldsymbol{\theta})$ denote the density function and $F(y|\boldsymbol{\theta}) = \Pr[Y \leq y]$ denote the cumulative distribution function of the discrete random variable, where $\boldsymbol{\theta}$ is a parameter vector. If realizations of $y$ less than a positive integer 1 are omitted, the ensuing zero-truncated density is given by

$$f(y|\boldsymbol{\theta},y \geq 1) = \frac{f(y|\boldsymbol{\theta})}{1-F(0|\boldsymbol{\theta})}, \qquad y = 1,2,.... \tag{2.8}$$

This specializes in the zero-truncated Poisson case, for example, to $f(y|\mu, y \geq 1) = e^{-\mu}\mu^y/[y!(1-\exp(-\mu))]$. It is straight-forward to construct a log-likelihood based on this density and to obtain maximum likelihood estimates.

*Censored* counts most commonly arise from aggregation of counts greater than some value. This is often done in survey design when the total probability mass over the aggregated values is relatively small. Censoring, like truncation, leads to inconsistent parameter estimates is the uncensored likelihood is mistakenly used.

For example, the number of events greater than some known value $c$ might be aggregated into a single category. Then some values of $y$ are incompletely observed; the precise value is unknown but it is known to equal or exceed $c$. The observed data has density

$$g(y|\boldsymbol{\theta}) = \begin{cases} f(y|\boldsymbol{\theta}) & \text{if } y < c, \\ 1-F(c|\boldsymbol{\theta}) & \text{if } y \geq c, \end{cases} \tag{2.9}$$

where $c$ is known. Specialization to the Poisson, for example, is straight-forward.

A related complication is that of *sample selection (*Terza, 1998). Then the count $y$ is observed only when another random variable, potentially correlated with $y$, crosses a threshold. For example, to see a medical specialist one may first need to see a general practitioner. Treatment of count data with sample selection is a current topic of research.

### 2.4. Overdispersion

The Poisson regression model is usually too restrictive for count data, leading to alternative models presented in Sections 3 and 4. The fundamental problem is that the distribution is parameterized in terms of a single scalar parameter ($\mu$) so that all moments of $y$ are a function of $\mu$. By contrast the normal distribution has separate parameters for location ($\mu$) and scale ($\sigma^2$). (For the same reason the one-parameter exponential is too restrictive for duration data and more general two-parameter distributions such as the Weibull are superior. Note that this complication does not arise with binary data. Then the distribution is clearly the one-parameter Bernoulli, as if the probability of success is $p$ then the probability of failure must be $1-p$. For binary data the issue is instead how to parameterize $p$ in terms of regressors.)

One way this restrictiveness manifests itself is that in many applications a Poisson density predicts the probability of a zero count to be considerably less than is actually observed in the sample. This is termed the *excess zeros* problem, as there are more zeros in the data than the Poisson predicts.

A second and more obvious way that the Poisson is deficient is that for count data the variance usually exceeds the mean, a feature called *overdispersion.* The Poisson instead implies equality of variance of mean, see (2.2), a property called *equidispersion.*

Overdispersion has qualitatively similar consequences to the failure of the assumption of homoskedasticity in the linear regression model. Provided the conditional mean is correctly specified, that is (2.3) holds, the Poisson MLE is still consistent. This is clear from inspection of the first-order conditions (2.5), since the left-hand side of (2.5) will have expected value of zero if $\mathsf{E}[y_i|\mathbf{x}_i] = \exp(\mathbf{x}_i'\boldsymbol{\beta})$. (This consistency property applies more generally to the quasi-MLE when the specified density is in the linear exponential family (LEF). Both Poisson and normal are members of the LEF.) It is nonetheless important to control for overdispersion for two reasons. First, in more complicated settings such as with truncation and censoring, overdispersion leads to the more fundamental problem of inconsistency. Second, even in the simplest settings large overdispersion leads to grossly deflated standard errors and grossly inflated $t$-statistics in the usual ML output.

A statistical test of overdispersion is therefore highly desirable after running a Poisson regression. Most count models with overdispersion specify overdispersion to be of the form

$$\mathsf{V}[y_i|\mathbf{x}_i] = \mu_i + \alpha g(\mu_i), \tag{2.10}$$

where $\alpha$ is an unknown parameter and $g(\cdot)$ is a known function, most commonly $g(\mu) = \mu^2$ or $g(\mu) = \mu$. It is assumed that under both null and alternative hypotheses the mean is correctly specified as, for example, $\exp(\mathbf{x}_i'\boldsymbol{\beta})$, while under the null hypothesis $\alpha = 0$ so that $\mathsf{V}[y_i|\mathbf{x}_i] = \mu_i$. A simple test statistic for $H_0 : \alpha = 0$ versus $H_1 : \alpha \neq 0$ or $H_1 : \alpha > 0$ can be computed by estimating the Poisson model, constructing fitted values $\widehat{\mu}_i = \exp(\mathbf{x}_i'\widehat{\boldsymbol{\beta}})$ and running the auxiliary OLS regression (without constant)

$$\frac{(y_i - \widehat{\mu}_i)^2 - y_i}{\widehat{\mu}_i} = \alpha \frac{g(\widehat{\mu}_i)}{\widehat{\mu}_i} + u_i, \tag{2.11}$$

where $u_i$ is an error term. The reported $t$-statistic for $\alpha$ is asymptotically normal under the null hypothesis of no overdispersion. This test can also be used for *underdispersion*, in which case the conditional variance is less than the conditional mean.

# 3. Other Parametric Count Regression Models

Various models that are less restrictive than Poisson are presented in this section.

First, overdispersion in count data may be due to unobserved heterogeneity. Then counts are viewed as being generated by a Poisson process, but the researcher is unable to correctly specify the rate parameter of this process. Instead the rate parameter is itself a random variable. This mixture approach, presented in Sections 3.1-3.2, leads to the widely-used negative binomial model.

Second, overdispersion, and in some cases underdispersion, may arise because the process generating the first event may differ from that determining later events. For example, an initial doctor consultation may be solely a patient's choice, while subsequent visits are also determined by the doctor. This leads to the hurdle model, presented in Section 3.3.

Third, overdispersion in count data may be due to failure of the assumption of independence of events which is implicit in the Poisson process. One can introduce dependence so that, for example, the occurrence of one doctor visit makes subsequent doctor visits more likely. This approach has not been widely used in count data analysis. (In duration data analysis this is called true state dependence, to be contrasted with the first approach of unobserved heterogeneity.) Particular assumptions again lead to the negative binomial; see also Winkelmann (1995). A discrete choice model that progressively models $\Pr[y = j|y \geq j - 1]$ is presented in Section 3.4, and issues of dependence also arise in Section 5 on time series.

Fourth, one can refer to the extensive and rich literature on univariate iid count distributions, which offers intriguing possibilities such as the logarithmic series and hypergeometric distribution (Johnson, Kotz, and Kemp, 1992). New regression models can be developed by letting one or more parameters be a specified function of regressors. Such models are not presented here. The approach has less motivation than the first three approaches and the resulting models may not be any better.

## 3.1. Continuous Mixture Models

The negative binomial model can be obtained in many different ways. The following justification using a mixture distribution is one of the oldest and has wide appeal.

Suppose the distribution of a random count $y$ is Poisson, conditional on the parameter $\lambda$, so that $f(y|\lambda) = \exp(-\lambda)\lambda^y/y!$. Suppose now that the parameter $\lambda$ is random, rather than being a completely deterministic function of regressors $\mathbf{x}$. In particular, let $\lambda = \mu\nu$, where $\mu$ is a deterministic function of $\mathbf{x}$, for example $\exp(\mathbf{x}'\boldsymbol{\beta})$, and $\nu > 0$ is iid distributed with density $g(\nu|\alpha)$. This is an example of *unobserved heterogeneity*, as different observations may have different $\lambda$ (heterogeneity) but part of this difference is due to a random (unobserved) component $\nu$.

The marginal density of $y$, unconditional on the random parameter $\nu$ but conditional on the deterministic parameters $\mu$ and $\alpha$, is obtained by integrating out $\nu$. This yields

$$h(y|\mu, \alpha) = \int f(y|\mu, \nu)g(\nu|\alpha)dv, \tag{3.1}$$

where $g(\nu|\alpha)$ is called the *mixing distribution* and $\alpha$ denotes the unknown parameter of the mixing distribution. The integration defines an "average" distribution. For some specific choices of $f(\cdot)$ and $g(\cdot)$, the integral will have an analytical or closed-form solution.

If $f(y|\lambda)$ is the Poisson density and $g(\nu)$, $\nu > 0$, is the gamma density with $\mathsf{E}[\nu] = 1$ and $\mathsf{V}[\nu] = \alpha$ we obtain the negative binomial density

$$h(y|\mu, \alpha) = \frac{\Gamma(\alpha^{-1} + y)}{\Gamma(\alpha^{-1})\Gamma(y+1)} \left(\frac{\alpha^{-1}}{\alpha^{-1} + \mu}\right)^{\alpha^{-1}} \left(\frac{\mu}{\mu + \alpha^{-1}}\right)^{y}, \qquad \alpha > 0, \tag{3.2}$$

where $\Gamma(\cdot)$ denotes the gamma integral which specializes to a factorial for an integer argument. Special cases of the negative binomial include the Poisson ($\alpha = 0$) and the geometric ($\alpha = 1$).

The first two moments of the negative binomial distribution are

$$\begin{aligned} \mathsf{E}[y|\mu, \alpha] &= \mu, \\ \mathsf{V}[y|\mu, \alpha] &= \mu(1 + \alpha\mu). \end{aligned} \tag{3.3}$$

The variance therefore exceeds the mean, since $\alpha > 0$ and $\mu > 0$. Indeed it can be shown easily that overdispersion always arises if $y|\lambda$ is Poisson and the mixing is of the form $\lambda = \mu\nu$ where $\mathsf{E}[\nu] = 1$. Note also that the overdispersion is of the form (2.10) discussed in Section 2.4.

Two standard variants of the negative binomial are used in regression applications. Both variants specify $\mu_i = \exp(\mathbf{x}_i'\boldsymbol{\beta})$. The most common variant lets $\alpha$ be a parameter to be estimated, in which case the conditional variance function, $\mu + \alpha\mu^2$ from (3.3), is quadratic in the mean. The log-likelihood is easily obtained from (3.2), and estimation is by maximum likelihood.

The other variant of the negative binomial model has a linear variance function, $\mathsf{V}[y|\mu, \alpha] = (1 + \delta)\mu$, obtained by replacing $\alpha$ by $\delta/\mu$ throughout (3.2). Estimation by ML is again straightforward. Sometimes this variant is called negative binomial 1 (NB1) in contrast to the variant with a quadratic variance function which has been called negative binomial 2 (NB2) model (Cameron and Trivedi, 1998).

The negative binomial model with quadratic variance function has been found to be very useful in applied work. It is the standard cross-section model for counts, which are usually overdispersed, along with the Quasi-MLE of section 4.1.

For mixtures other than Poisson-gamma, such as those that instead use as mixing distribution the lognormal distribution or the inverse-Gaussian distribution, the marginal distribution cannot be expressed in a closed form. Then one may have to use numerical quadrature or simulated maximum likelihood to estimate the model. These methods are entirely feasible with currently available computing power. If one is prepared to use simulation-based estimation methods, see Gourieroux and Monfort (1997), the scope for using mixed-Poisson models of various types is very extensive.

### 3.2. Finite Mixture Models

The mixture model in the previous subsection was a continuous mixture model, as the mixing random variable $\nu$ was assumed to have continuous distribution. An alternative approach instead uses a *discrete* representation of unobserved heterogeneity, which generates a class of models called *finite mixture models.* This class of models is a particular subclass of *latent class models.*

In empirical work the more commonly used alternative to the continuous mixture is in the class of modified count models discussed in the next section. However, it is more natural to follow up the preceding section with a discussion of finite mixtures. Further, the subclass of modified count models can be viewed as a special case of finite mixtures.

We suppose that the density of $y$ is a linear combination of $m$ different densities, where the $j^{th}$ density is $f_j(y|\lambda_j)$, $j = 1, 2, ..., m$. Thus an $m$-component finite mixture is

$$f(y|\boldsymbol{\lambda}, \boldsymbol{\pi}) = \sum_{j=1}^{m} \pi_j f_j(y|\lambda_j), \qquad 0 < \pi_j < 1, \ \sum_{j=1}^{m} \pi_j = 1. \tag{3.4}$$

For example, in a study of the use of medical services with $m = 2$, the first density may correspond to heavy users of the service and the second to relatively low users, and the fractions of the two types in the populations are $\pi_1$ and $\pi_2 (= 1 - \pi_1)$ respectively.

The goal of the researcher who uses this model is to estimate the unknown parameters $\lambda_j$, $j = 1, ..., m$. It is easy to develop regression models based on (3.4). For example, if NB2 models are used then $f_j(y|\lambda_j)$ is the NB2 density (3.2) with parameters $\mu_j = \exp(\mathbf{x}'\boldsymbol{\beta}_j)$ and $\alpha_j$, so $\lambda_j = (\boldsymbol{\beta}_j, \alpha_j)$. If the number of components, $m$, is given, then under some regularity conditions maximum likelihood estimation of the parameters $(\pi_j, \lambda_j)$, $j = 1, ..., m$, is possible. The details of the estimation methods, less straightforward due to the presence of the mixing parameters $\pi_j$, is omitted here because of space constraints. See Cameron and Trivedi (1998, Chapter 4). It is possible also to probabilistically assign each case to a subpopulation (in the sense that the estimated probability of the case belonging to that subpopulation is the highest) *after* the model has been estimated.

### 3.3. Modified Count Models

The leading motivation for modified count models is to solve the so-called problem of excess zeros, the presence of more zeros in the data than predicted by count models such as the Poisson.

The *hurdle model* or *two-part model* relaxes the assumption that the zeros and the positives come from the same data generating process. The zeros are determined by the density $f_1(\cdot)$, so that $\Pr[y = 0] = f_1(0)$. The positive counts come from the truncated density $f_2(y|y > 0) = f_2(y)/(1 - f_2(0))$, which is multiplied by $\Pr[y > 0] = 1 - f_1(0)$ to ensure that probabilities sum to unity. Thus

$$g(y) = \begin{cases} f_1(0) & \text{if } y = 0, \\ \dfrac{1 - f_1(0)}{1 - f_2(0)} f_2(y) & \text{if } y \geq 1. \end{cases} \tag{3.5}$$

This reduces to the standard model only if $f_1(\cdot) = f_2(\cdot)$. Thus in the modified model the two processes generating the zeros and the positives are not constrained to be the same. While the motivation for this model is to handle excess zeros, it is also capable of modeling too few zeros.

Maximum likelihood estimation of the hurdle model involves separate maximization of the two terms in the likelihood, one corresponding to the zeros and the other to the positives. This is straight-forward.

A hurdle model has the interpretation that it reflects a two-stage decision-making process. For example, a patient may initiate the first visit to a doctor, but the second and subsequent visits may be determined by a different mechanism (Pohlmeier and Ulrich, 1995).

Regression applications use hurdle versions of the Poisson or negative binomial, obtained by specifying $f_1(\cdot)$ and $f_2(\cdot)$ to be the Poisson or negative binomial densities given earlier. In application the covariates in the hurdle part which models the zero/one outcome need not be the same as those which appear in the truncated part, although in practice they are often the same. The hurdle model is widely used, and the hurdle negative binomial model is quite flexible. Drawbacks are that the model is not very parsimonious, typically the number of parameters is doubled, and parameter interpretation is not as easy as in the same model without hurdle.

The conditional mean in the hurdle model is the product of probability of positives and the conditional mean of the zero-truncated density. Therefore, using a Poisson regression when the hurdle model is the correct specification implies a misspecification which will lead to inconsistent estimates.

### 3.4. Discrete Choice Models

Count data can be modelled by discrete choice model methods, possibly after some grouping of counts to limit the number of categories. For example the categories may be 0, 1, 2, 3 and 4 or more if few observations exceed four. Unordered models such as multinomial logit are not parsimonious and more importantly are inappropriate. Instead one should use a sequential discrete choice model that recognizes the ordering of the data, such as ordered logit or ordered probit.

## 4. Partially Parametric Models

By partially parametric models we mean that we focus on modeling the data via the conditional mean and variance, and even these may not be fully specified. In Section 5.1 we consider models based on specification of the conditional mean and variance. In Section 5.2 we consider and critique the use of least squares methods that do not explicitly model the heteroskedasticity inherent in count data. In Section 5.3 we consider models that are even more partially parametric, such as incomplete specification of the conditional mean.

### 4.1. Quasi-ML Estimation

In the econometric literature *pseudo-ML* (PML) or *quasi-ML* (QML) estimation refers to estimating by ML, under the assumption that the specified density is correct (Gourieroux et al. 1984a). PML and QML are often used interchangeably. The distribution of the estimator is obtained under weaker assumptions about the data generating process than those that led to the specified likelihood function. In the statistics literature QML often refers to nonlinear generalized least squares estimation. For the Poisson regression QML in the latter sense is equivalent to standard maximum likelihood.

From (2.5), the Poisson PML estimator, $\widehat{\boldsymbol{\beta}}_P$, has first-order conditions $\sum_{i=1}^{n}(y_i - \exp(\mathbf{x}_i'\boldsymbol{\beta}))\mathbf{x}_i = \mathbf{0}$. As already noted in Section 2.4, the summation on the left-hand side has expectation zero if $\mathsf{E}[y_i|\mathbf{x}_i] = \exp(\mathbf{x}_i'\boldsymbol{\beta})$. Hence the Poisson PML is consistent under the weaker assumption of correct specification of the conditional mean – the data need not be Poisson distributed. Using standard results, the variance matrix is of the sandwich form, with

$$\mathsf{V}_{PML}[\widehat{\boldsymbol{\beta}}_{\mathsf{P}}] = \left(\sum_{i=1}^{n} \mu_i \mathbf{x}_i \mathbf{x}_i'\right)^{-1} \left(\sum_{i=1}^{n} \omega_i \mathbf{x}_i \mathbf{x}_i'\right) \left(\sum_{i=1}^{n} \mu_i \mathbf{x}_i \mathbf{x}_i'\right)^{-1} \tag{4.1}$$

and $\omega_i = \mathsf{V}[y_i|\mathbf{x}_i]$ is the conditional variance of $y_i$.

Given an assumption for the functional form for $\omega_i$, and a consistent estimate $\widehat{\omega}_i$ of $\omega_i$, one can consistently estimate this covariance matrix. We could use the Poisson assumption, $\omega_i = \mu_i$, but as already noted the data are often overdispersed, with $\omega_i > \mu_i$. Common variance functions used are $\omega_i = (1+\alpha\mu_i)\mu_i$, that of the NB2 model discussed in Section 3.1, and $\omega_i = (1+\alpha)\mu_i$, that of the NB1 model. Note that in the latter case (4.1) simplifies to $\mathsf{V}_{PML}[\widehat{\boldsymbol{\beta}}_{\mathsf{P}}] = (1+\alpha)\left(\sum_{i=1}^{n} \mu_i \mathbf{x}_i \mathbf{x}_i'\right)^{-1}$, so with overdispersion ($\alpha > 0$) the usual ML variance matrix given in (2.6) is understating the true variance.

If $\omega_i = \mathsf{E}[(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2|\mathbf{x}_i]$ is instead unspecified, a consistent estimate of $\mathsf{V}_{PML}[\widehat{\boldsymbol{\beta}}_{\mathsf{P}}]$ can be obtained by adapting the Eicker-White robust sandwich variance estimate formula to this case. The middle sum in (4.1) needs to be estimated. If $\widehat{\mu}_i \xrightarrow{p} \mu_i$ then $n^{-1}\sum_{i=1}^{n}(y_i - \widehat{\mu}_i)^2 \mathbf{x}_i \mathbf{x}_i' \xrightarrow{p} \lim n^{-1}\sum_{i=1}^{n} \omega_i \mathbf{x}_i \mathbf{x}_i'$. Thus a consistent estimate of $\mathsf{V}_{PML}[\widehat{\boldsymbol{\beta}}_P]$ is given by (4.1) with $\omega_i$ and $\mu_i$ replaced by $(y_i - \widehat{\mu}_i)^2$ and $\widehat{\mu}_i$.

When doubt exists about the form of the variance function, the use of the PML estimator is recommended. Computationally this is essentially the same as Poisson ML, with the qualification that the variance matrix must be recomputed. The calculation of robust variances is often an option in standard packages.

These results for Poisson PML estimation are qualitatively similar to those for PML estimation in the linear model under normality. They extend more generally to PML estimation based on densities in the linear exponential family. In all cases consistency requires only correct specification of the conditional mean (Nelder and Wedderburn, 1972; Gourieroux et al., 1984a). This has led to a vast statistical literature on *generalized linear models* (GLM), see McCullagh and Nelder (1989), which permits valid inference providing the conditional mean is correctly specified and nests many types of data as special cases – continuous (normal), count (Poisson), discrete (binomial) and positive (gamma). Many methods for complications, such as time series and panel data models, are presented in the more general GLM framework rather than specifically for count data.

Many econometricians find it more natural to use the *generalized methods of moments* (GMM) framework rather than GLM. Then the starting point is the conditional moment $\mathsf{E}[y_i - \exp(\mathbf{x}_i'\boldsymbol{\beta})|\mathbf{x}_i] = 0$. If data are independent over $i$ and the conditional variance is a multiple of the mean it can be shown that the optimal choice of instrument is $\mathbf{x}_i$, leading to the estimating equations (2.5); for more detail, see Cameron and Trivedi (1998, 37-44). The GMM framework has been fruitful for panel data on counts, see Section 5.3, and for *endogenous* regressors. Fully specified simultaneous equations models for counts have not been yet developed, so instrumental variables methods are used. Given instruments $\mathbf{z}_i$,

$\dim(\mathbf{z}) \geq \dim(\mathbf{x})$, satisfying $\mathsf{E}[y_i - \exp(\mathbf{x}_i'\boldsymbol{\beta})|\mathbf{z}_i] = 0$, a consistent estimator of $\boldsymbol{\beta}$ minimizes

$$Q(\boldsymbol{\beta}) = \left(\sum_{i=1}^{n}(y_i - \exp(\mathbf{x}_i'\boldsymbol{\beta}))\mathbf{z}_i\right)' \mathbf{W} \left(\sum_{i=1}^{n}(y_i - \exp(\mathbf{x}_i'\boldsymbol{\beta}))\mathbf{z}_i\right),$$

where $\mathbf{W}$ is a symmetric weighting matrix.

### 4.2. Least Squares Estimation

When attention is focused on modeling just the conditional mean, least squares methods are inferior to the approach of the previous subsection.

Linear least squares regression of $y$ on $\mathbf{x}$ leads to consistent parameter estimates if the conditional mean is linear in $\mathbf{x}$. But for count data the specification $\mathsf{E}[y|\mathbf{x}] = \mathbf{x}'\boldsymbol{\beta}$ is inadequate as it permits negative values of $\mathsf{E}[y|\mathbf{x}]$. For similar reasons the linear probability model is inadequate for binary data.

Transformations of $y$ may be considered. In particular the logarithmic transformation regresses $\ln y$ on $\mathbf{x}$. This transformation is problematic if the data contain zeros, as is often the case. One standard solution is to add a constant term, such as 0.5, and to model $\ln(y + .5)$ by OLS. This method often produces unsatisfactory results, and complicates the interpretation of coefficients. It is also unnecessary as software to estimate basic count models is widely available.

### 4.3. Semiparametric Models

By *semiparametric models* we mean partially parametric models that have an infinite-dimensional component.

One example is optimal estimation of the regression parameters $\boldsymbol{\beta}$, when $\mu_i = \exp(\mathbf{x}_i'\boldsymbol{\beta})$ is assumed but $\mathsf{V}[y_i|\mathbf{x}_i] = \omega_i$ is left unspecified. The infinite-dimensional component arises because as $n \to \infty$ there are infinitely many variance parameters $\omega_i$. An optimal estimator of $\boldsymbol{\beta}$, called an *adaptive estimator*, is one that is as efficient as that when $\omega_i$ is known. Delgado and Kniesner (1997) extend results for the linear regression model to count data with exponential conditional mean function, using kernel regression methods to estimate weights to be used in a second-stage nonlinear least squares regression. In their application the estimator shows little gain over specifying $\omega_i = \mu_i(1 + \alpha\mu_i)$, overdispersion of the NB2 form.

A second class of semiparametric models incompletely specifies the conditional mean. Leading examples are *single-index models* and *partially linear models.* Single-index models specify $\mu_i = g(\mathbf{x}_i'\boldsymbol{\beta})$ where the functional form $g(\cdot)$ is left unspecified. Partially linear models specify $\mu_i = \exp(\mathbf{x}_i'\boldsymbol{\beta}+g(\mathbf{z}_i))$ where the functional form $g(\cdot)$ is left unspecified. In both cases root$-n$ consistent asymptotically normal estimators of $\boldsymbol{\beta}$ can be obtained, without knowledge of $g(\cdot)$.

## 5. Time Series, Multivariate and Panel Data

In this section we very briefly present extension from cross-section to other types of count data (see Cameron and Trivedi, 1998, for further detail). For time series and multivariate count data many models have been proposed but preferred methods have not yet been established. For panel data there is more agreement in the econometrics literature on which methods to use, though a wider range of models is considered in the statistics literature.

### 5.1. Time Series Data

If a time series of count data is generated by a Poisson point process then event occurrences in successive time intervals are independent. Independence is a reasonable assumption when the underlying stochastic process for events, conditional on covariates, has no memory. Then there is no need for special time series models. For example, the number of deaths (or births) in a region may be uncorrelated over time. At the same time the population, which cumulates births and deaths, will be very highly correlated over time.

The first step for time series count data is therefore to test for serial correlation. A simple test first estimates a count regression such as Poisson, obtains the residual, usually $(y_t - \exp(x_t'\widehat{\boldsymbol{\beta}}))$ where $x_t$ may include time trends, and tests for zero correlation between current and lagged residuals, allowing for the complication that the residuals will certainly be heteroskedastic.

Upon establishing the data are indeed serially correlated, there are several models to choose from. An esthetically appealing model is the INAR(1) model (*integer autoregressive model* of order one (INAR(1)) and its generalization to the negative binomial and to higher orders of serial correlation. This model specifies $y_t = \rho_t \circ y_{t-1} + \varepsilon_t$, where $\rho_t$ is a correlation parameter with $0 < \rho_t < 1$, for example $\rho_t = 1/[1 + \exp(-\mathbf{z}_t'\gamma)]$. The symbol $\circ$ denotes the *binomial thinning* operator, whereby $\rho_t \circ y_{t-1}$ is the realized value of a binomial random variable with probability of success $\rho_t$ in each of $y_{t-1}$ trials. One may think of each event as having a replication or survival probability of $\rho_t$ in the following period. As in a linear first order Markov model, this probability decays geometrically. A Poisson INAR(1) model, with a Poisson marginal distribution for $y_t$ arises when $\varepsilon_t$ is Poisson distributed with mean, say, $\exp(\mathbf{x}_t'\boldsymbol{\beta})$. A negative binomial INAR(1) model arises if $\varepsilon_t$ is negative binomial distributed.

An *autoregressive model*, or *Markov model*, is a simple adjustment to the earlier cross-section count models that directly enters lagged values of $y$ into the formula for the conditional mean of current $y$. For example, we might suppose $y_t$ conditional on current and past $\mathbf{x}_t$ and past $y_t$ is Poisson distributed with mean $\exp(\mathbf{x}_t'\boldsymbol{\beta}+\rho \ln y_{t-1}^*)$, where $y_{t-1}^*$ is an adjustment to ensure a non-zero lagged value, such as $y_{t-1}^* = \ln(y_{t-1}+0.5)$ or $y_{t-1}^* = \max(0.5, y_{t-1})$.

*Serially correlated error models* induce time series correlation by introducing unobserved heterogeneity, see Section 3.1, and allowing this to be serially correlated. For example, $y_t$ is Poisson distributed with mean $\exp(\mathbf{x}_t'\boldsymbol{\beta})\nu_t$ where $\nu_t$ is a serially correlated random variable, (Zeger, 1988).

*State space models* or *time-varying parameters models* allow the conditional mean to be random variable drawn from a distribution whose parameters evolve over time. For example, $y_t$ is Poisson distributed with mean $\mu_t$ where $\mu_t$ is a draw from a gamma distribution, (Harvey and Fernandes, 1989).

*Hidden Markov models* specify different parametric models in different regimes, and induce serial correlation by specifying the stochastic process determining which regime currently applies to be an unobserved Markov process (MacDonald and Zucchini, 1997).

### 5.2. Multivariate Data

In some data sets more than one count is observed. For example, data on the utilization of several different types of health service, such as doctor visits and hospital days, may be available. Joint modeling will improve efficiency and provide richer models of the data if counts are correlated.

Most parametric studies have used the *bivariate Poisson.* This model, however, is too restrictive as it implies variance-mean equality for the counts and restricts the correlation to be positive. Development of better parametric models is a current area of research.

### 5.3. Panel Data

One of the major and earliest applications of count data methods in econometrics is to panel data on the number of patents awarded to firms over time (Hausman, Hall, and Griliches, 1984). The starting point is the Poisson regression model with exponential conditional mean and multiplicative individual-specific term

$$y_{it} \sim \mathsf{P}[\alpha_i \exp(\mathbf{x}_{it}'\boldsymbol{\beta})], \qquad i = 1, ..., n, \quad t = 1, ..., T, \tag{5.1}$$

where we consider a short panel with $T$ small and $n \to \infty$. As in the linear case, both fixed effects and random effects models are possible.

The *fixed effects model* lets $\alpha_i$ be an unknown parameter. This parameter can be eliminated by quasi-differencing and modeling the transformed random variable $y_{it} - (\lambda_{it}/\bar{\lambda}_i)\bar{y}_i$, where $\bar{\lambda}_i$ and $\bar{y}_i$ denote the individual-specific means of $\lambda_{it}$ and $y_{it}$. By construction this has zero mean, conditional on $\mathbf{x}_{i1}, ..., \mathbf{x}_{iT}$. A moments-based estimator of $\boldsymbol{\beta}$ then solves the sample moment condition $\sum_{i=1}^{n} \sum_{t=1}^{T} \mathbf{x}_{it}(y_{it} - (\lambda_{it}/\bar{\lambda}_i)\bar{y}_i) = \mathbf{0}$.

An alternative to the quasi-differencing approach is the conditional likelihood approach that was followed By Hausman et al. (1984). In this approach the fixed effects are eliminated by conditioning the distribution of counts on $\sum_{t=1}^{T} y_{it}$.

The *random effects model* lets $\alpha_i$ be a random variable with specified distribution that depends on parameters, say $\boldsymbol{\delta}$. The random effects are integrated out, in a similar way to the unobserved heterogeneity in Section 3.1, and the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ are estimated by maximum likelihood. In some cases, notably when $\alpha_i$ is gamma distributed, a closed form solution is obtained upon integrating out $\alpha_i$. In other cases, such as normally distributed random effects, a closed form solution is not obtained, but ML estimation based on numerical integration is feasible.

*Dynamic panel data models* permit the regressors $\mathbf{x}$ to include lagged values of $y$. Several studies use the fixed effects variant of (5.1), where $\mathbf{x}_{it}$ now includes, for example, $y_{it-1}$. This is an autoregressive count model, see Section 5.1, adapted to panel data. The quasi-differencing procedure for the non-dynamic fixed effects case can be adapted to the dynamic case.

## 6. Practical Considerations

Those with experience of nonlinear least squares will find it easy to use packaged software for Poisson regression, which is a widely available option in popular econometrics packages like

LIMDEP, STATA and TSP. One should ensure, however, that reported standard errors are based on (4.1) rather than (2.6). Many econometrics packages also include negative binomial regression, also widely-used for cross-section count regression, and the basic panel data models. Statistics packages such as SAS and SPSS include count regression in a generalized linear models module. Standard packages also produce some goodness-of-fit statistics, such as the $G^2$-statistic and pseudo-$R^2$ measures, for the Poisson (see Cameron and Windmeijer, 1996).

More recently developed models, such as finite mixture models, most time series models and dynamic panel data models, require developing one's own programs. A promising route is to use matrix programming languages such as GAUSS, MATLAB, SAS/IML or SPLUS in conjunction with software for implementing estimation based on user-defined objective functions. For simple models packages such as LIMDEP, STATA and TSP make it possible to implement maximum likelihood estimation and (highly desirable) robust variance estimation for user-defined functions.

In addition to reporting parameter estimates it is useful to have an indication of the magnitude of the estimated effects, as discussed in Section 2.2. And as noted in Section 2.4, care should be taken to ensure that reported standard errors and t-statistics for the Poisson regression model are based on variance estimates robust to overdispersion.

In addition to estimation it is strongly recommended that specification tests are used to assess the adequacy of the estimated model. For Poisson cross-section regression overdispersion tests are easy to implement. For time series regression tests of serial correlation should be used. For any parametric model one can compare the actual and fitted frequency distribution of counts. Formal statistical specification and goodness-of-fit tests based on actual and fitted frequencies are available.

In most practical situations one is likely to face the problem of model selection. For likelihood-based models that are nonnested one can use selection criteria, such as the Akaike and Schwarz criteria, which are based on the fitted log-likelihood but with degrees of freedom penalty for models with many parameters.

## 7. Further reading

All the topics dealt with in this chapter are treated at greater length and depth in Cameron and Trivedi (1998) which also provides a comprehensive bibliography. Winkelmann (1997) also provides a fairly complete treatment of the econometric literature on counts. The statistics literature generally analyzes counts in the context of generalized linear models (GLM). The standard reference is McCullagh and Nelder (1989). The econometrics literature generally fails to appreciate the contributions of the GLM literature on generalized linear models. Fahrmeier and Tutz (1994) provide a recent and more econometric exposition of GLMs.

The material in Section 2 is very standard and appears in many places. A similar observation applies to the negative binomial model in section Section 3.1. Cameron and Trivedi (1986) provide an early presentation and application. For the finite mixture approach of Section 3.2 see Deb and Trivedi (1997). Applications of the hurdle model in Section 3.3 include Mullahy (1986), who first proposed the model, Pohlmeier and Ulrich (1995), and Gurmu and Trivedi (1996). The quasi-MLE of section 4.1 is presented in detail by Gourieroux et al. (1984a, 1984b) and by Cameron and Trivedi (1996).

Regression models for the types of data discussed in Section 5 are in their infancy. The notable exception is that (static) panel data count models are well established, with the standard reference being Hausman et al. (1984). See also Brannas and Johansson (1996). For reviews of the various time series models see MacDonald and Zucchini (1997, chapter 2) and Cameron and Trivedi (1998, chapter 7). Developing adequate regression models for multivariate count data is currently an active area. For dynamic count data models there are several recent references, including Blundell et al. (1995)

For further discussion of diagnostic testing, only briefly mentioned in Section 6, see Cameron and Trivedi (1998, chapter 5).

# References

Blundell, R., R. Griffith, and J. Van Reenen (1995) “Dynamic Count Data Models of Technological Innovation”, *Economic Journal*, 105, 333-44.

Brännäs, K. and P. Johansson (1996), “Panel Data Regression for Counts,” *Statistical Papers*, 37, 191-213.

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

Cameron, A.C., P.K. Trivedi, F. Milne and J. Piggott (1988), “A Microeconometric Model of the Demand for Health Care and Health Insurance in Australia”, *Review of Economic Studies*, 55, 85-106.

Cameron, A.C. and F.A.G. Windmeijer (1996), “R-Squared Measures for Count Data Regression Models with Applications to Health Care Utilization”, *Journal of Business and Economic Statistics*, 14, 209-220.

Davutyan, N. (1989), “Bank Failures as Poisson Variates”, *Economic Letters,* 29, 333-338.

Dean, C. and R. Balshaw (1997), “Efficiency Lost by Analyzing Counts Rather than Event Times in Poisson and Overdispersed Poisson Regression Models ”, *Journal of the American Statistical Association*, 92, 1387-1398.

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.

Delgado, M.A. and T.J. Kniesner (1997), “Count Data Models with Variance of Unknown Form: An Application to a Hedonic Model of Worker Absenteeism,” *Review of Economics and Statistics*, 79, 41-49.

Fahrmeier, L. and G.T. Tutz (1994), *Multivariate Statistical Modelling Based on Generalized Linear Models*, New York: Springer-Verlag.

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

Gourieroux, C., A. Monfort and A. Trognon (1984a), "Pseudo Maximum Likelihood Methods: Theory", *Econometrica*, 52, 681-700.

Gourieroux, C., A. Monfort and A. Trognon (1984b), "Pseudo Maximum Likelihood Methods: Applications to Poisson Models", *Econometrica*, 52, 701-720.

Gurmu, S. and P.K. Trivedi (1996), "Excess Zeros in Count Models for Recreational Trips", *Journal of Business and Economic Statistics*, 14, 469-477.

Harvey, A.C. and C. Fernandes (1989), "Time Series Models for Count or Qualitative Observations (with Discussion)", *Journal of Business and Economic Statistics*, 7, 407-417.

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.

Johnson, N. L., S. Kotz and A.W. Kemp (1992), *Univariate Distributions,* Second edition, New York: John Wiley.

MacDonald, I.L. and W. Zucchini (1997), *Hidden Markov and other Models for Discrete-valued Time Series,* London: Chapman and Hall.

McCullagh, P. and J.A. Nelder (1989), *Generalized Linear Models*, Second Edition, London: Chapman and Hall.

Mullahy, J. (1986), "Specification and Testing of Some Modified Count Data Models," *Journal of Econometrics*, 33, 341-365.

Nelder, J.A. and R.W.M. Wedderburn (1972), "Generalized Linear Models", *Journal of the Royal Statistical Society* A, 135, 370-384.

Patil, G.P. (1970), editor, *Random Counts in Models and Structures*, Volumes 1-3, University Park and London: Pennsylvania State University Press.

Pohlmeier,W. and V.Ulrich (1995), "An Econometric Model of the Two-Part Decision-making Process in the Demand for Health Care", *Journal of Human Resources*, 30, 339-361.

Rose, N. (1990), "Profitability and Product Quality: Economic Determinants of Airline Safety Performance", *Journal of Political Economy*, 98, 944-964.

Terza, J. (1998), "Estimating Count Data Models with Endogenous Switching: Sample Selection and Endogenous Switching Effects", *Journal of Econometrics*, 84, 129-139.

Winkelmann, R. (1995), "Duration Dependence and Dispersion in Count-Data Models," *Journal of Business and Economic Statistics*, 13, 467-474.

Winkelmann, R. (1997), *Econometric Analysis of Count Data*, Berlin, Springer-Verlag.

Zeger, S.L. (1988), “A Regression Model for Time Series of Counts”, *Biometrika*, 75, 621-629.