JIRSS (2002)
Vol. 1, Nos. 1-2, pp 7-33

Modeling Nonnegative Data with Clumping at Zero: A Survey

Yongyi Min, Alan Agresti

Department of Statistics, University of Florida, Gainesville, Florida, USA 32611-8545. (ymin@stat.ufl.edu, aa@stat.ufl.edu)

Abstract. Applications in which data take nonnegative values but have a substantial proportion of values at zero occur in many disciplines. The modeling of such "clumped-at-zero" or "zero-inflated" data is challenging. We survey models that have been proposed. We consider cases in which the response for the non-zero observations is continuous and in which it is discrete. For the continuous and then the discrete case, we review models for analyzing cross-sectional data. We then summarize extensions for repeated measurement analyses (e.g., in longitudinal studies), for which the literature is still sparse. We also mention applications in which more than one clump can occur and we suggest problems for future research.

1 Introduction

In some applications, the response variable can take any nonnegative value but has positive probability of a zero outcome. We refer to a variable as *semicontinuous* when it has a continuous distribution

Received: May 2002

Key words and phrases: Compliance, finite mixture model, logistic regression, Neyman type a distribution, proportional odds model, semicontinuous data, Tobit model, zero-inflated data.

except for a probability mass at 0. Semicontinuous data are common in many areas. For example, when each observation is a record of the total rainfall in the previous day, many days have no rainfall. In a study of household expenditures, some households spend nothing on a certain commodity during the period of investigation. In a study of annual medical costs, a portion of the population has zero medical expense. With semicontinuous data, unlike left-censored data, the zeros represent actual response outcomes.

A related type of data are zero-inflated count data. These are data that have a higher proportion of zeros than expected under standard distributional assumptions such as the Poisson. Such data are also common in a variety of disciplines. Examples of variables that one might expect to be zero-inflated are observations for the past month of the reported number of times participating in sports activities, the number of times one has visited a doctor, and the frequency of sexual intercourse.

One difficulty with semicontinuous data analysis is that the existence of a probability mass at zero makes common response distributions such as the normal or gamma inappropriate for modeling the data. Likewise for zero-inflated count data, a generalized linear model based on Poisson or overdispersed count distributions usually encounters lack of fit due to disproportionately large frequencies of zeros. Thus, these types of data stimulate interesting modeling problems. Some statistical methodology has been developed to deal with them. This article surveys methods that have been proposed for modeling these two types of data that have clumping at 0.

Section 2 introduces models for semicontinuous data and then summarizes their advantages and disadvantages. Section 3 introduces models for zero-inflated count data. Section 4 surveys extensions of these two types of models to handle repeated measurement, such as in longitudinal studies. Section 5 discusses a data type that has a clump at both boundaries of a sample space, such as occurs with the medical application of studying subjects' compliance in taking prescribed drugs. The final section suggests possible areas for future research.

2 Models for Semicontinuous Data

This section introduces some methods for modeling semicontinuous data. The early research on modeling such data appeared mainly in

the econometrics literature. Tobin (1958) proposed a censored regression model to describe household expenditures on durable goods. This model is now commonly referred to as the *Tobit model*. The term "Tobit" arose from its similarities in derivation to the probit model, based on a normal latent variable construction described below. Since then, related literature contains numerous econometric applications as well as various generalizations of the Tobit model (e.g., Cragg 1971, Amemiya 1973, Gronau 1974, Heckman 1974, 1979). These all posit an underlying normal random variable that is censored by a random mechanism.

An alternative strand of literature for semicontinuous data does not assume an underlying normal distribution. Duan, Manning, Morris, and Newhouse (1983) proposed a two-part model to fit data on expenditures for medical care. Jørgensen (1987) proposed a compound Poisson exponential dispersion model for semicontinuous data. Saei, Ward, and McGilchrist (1996) applied an ordinal response model that requires grouping the response outcomes into categories. The Tobit model and these alternative models are described in the following subsections.

2.1 Tobit models

For response variable Y, let y_i denote the observation for subject i, $i = 1, \ldots, n$. The Tobit model assumes an underlying normally distributed variable Y_i^* such that:

$$y_i = \begin{cases} y_i^*, & \text{if } y_i^* > 0 \\ 0, & \text{if } y_i^* \leq 0 \end{cases}$$

When $y_i^* \leq 0$, its value is unobserved.

Including explanatory variables, the model assumes that the underlying variable is generated by

$$y_i^* = \boldsymbol{x}_i'\boldsymbol{\beta} + u_i$$

where $\boldsymbol{x}_i$ is a column vector of explanatory variable values for subject i and $\{u_i\}$ are independent from a normal $N(0, \sigma^2)$ distribution. Let $\Phi(\cdot)$ and $\phi(\cdot)$ denote the cumulative distribution function (*cdf*) and the probability density function (*pdf*) of the $N(0, 1)$ distribution. For

the Tobit model, the probability of a zero response is

$$\begin{aligned} P(Y_i = 0) = P(\boldsymbol{x}_i'\boldsymbol{\beta} + u_i \leq 0) &= P(u_i \leq -\boldsymbol{x}_i'\boldsymbol{\beta}) \\ &= \Phi\left(\frac{-\boldsymbol{x}_i'\boldsymbol{\beta}}{\sigma}\right) \\ &= 1 - \Phi\left(\frac{\boldsymbol{x}_i'\boldsymbol{\beta}}{\sigma}\right) \end{aligned}$$

Conditional on $y_i > 0$, its probability density function is

$$f(y_i; \boldsymbol{\beta}, \sigma) = \sigma^{-1}\phi\left(\frac{y_i - \boldsymbol{x}_i'\boldsymbol{\beta}}{\sigma}\right)$$

Thus, the likelihood function for a sample of n independent observations is

$$\ell(\boldsymbol{\beta}, \sigma) = \left[\prod_{y_i=0}\{1 - \Phi\left(\frac{\boldsymbol{x}_i'\boldsymbol{\beta}}{\sigma}\right)\}\right]\left[\prod_{y_i>0}\sigma^{-1}\phi\left(\frac{y_i - \boldsymbol{x}_i'\boldsymbol{\beta}}{\sigma}\right)\right]$$

Tobin (1958) used a Newton-Raphson algorithm to find the maximum likelihood (ML) estimates of $\boldsymbol{\beta}$ and σ. Amemiya (1984) presented a comprehensive survey of the Tobit model and its generalizations.[1]

The Tobit model assumes normality for the distribution of the error term, with constant variance. In many applications this is unrealistic. When the model form is correct but the distribution of u_i is not normal, the ML estimators are inconsistent (Robinson 1982).

Powell (1986) proposed semi-parametric estimation for the Tobit model. He used a symmetrically trimmed least squares (STLS) estimator. This assumes that $\{u_i\}$ are symmetrically distributed about zero. The STLS estimator is defined as

$$\hat{\boldsymbol{\beta}}_{STLS} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} I(\boldsymbol{x}_i'\boldsymbol{\beta} > 0)[\min(y_i, 2\boldsymbol{x}_i'\boldsymbol{\beta}) - \boldsymbol{x}_i'\boldsymbol{\beta}]^2$$

where I is the indicator function. For a given $\boldsymbol{\beta}$, the sum in this expression deletes the observations with $\boldsymbol{x}_i'\boldsymbol{\beta} \leq 0$. When $\boldsymbol{x}_i'\boldsymbol{\beta} > 0$, the lower tail of the distribution of Y_i is censored at zero; symmetrically censoring the upper tail of the distribution (essentially by replacing y_i by $\min\{y_i, 2\boldsymbol{x}_i'\boldsymbol{\beta}\}$) restores the symmetry of distribution of Y^*. The resulting estimator $\hat{\boldsymbol{\beta}}_{STLS}$ is consistent and asymptotically normal

[1] James Tobin, Sterling Professor Emeritus of Economics at Yale University, won the 1981 Nobel Prize in Economics; he died on March 11, 2002.

under the symmetrical distribution assumption (Powell 1986). An iterative procedure yields $\hat{\boldsymbol{\beta}}_{STLS}$.

Yoo, Kim, and Lee (2001) used this method with the bootstrap to estimate the covariance matrix of $\hat{\boldsymbol{\beta}}_{STLS}$. For M bootstrap replications with estimate $\hat{\boldsymbol{\beta}}_j$ in replication j, their estimate is

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{M}\sum_{j=1}^{M}(\hat{\boldsymbol{\beta}}_j - \bar{\boldsymbol{\beta}}_{STLS})(\hat{\boldsymbol{\beta}}_j - \bar{\boldsymbol{\beta}}_{STLS})'$$

where $\bar{\boldsymbol{\beta}}_{STLS} = (1/M)\sum_{j=1}^{M}\hat{\boldsymbol{\beta}}_j$. In an empirical study, Yoo et al. showed that semi-parametric estimation significantly outperforms estimation assuming normality (i.e., the Tobit model).

2.2 Two-part models

The Tobit model allows the same underlying stochastic process to determine whether the response is zero or positive as well as the value of a positive response. That is, the same parameters influence whether the outcome is zero or positive as well as the magnitude of the outcome, conditional on its being positive. The next two subsections discuss "two-part models" that allow the two components to have different parameters.

Without assuming an underlying normal distribution, Duan et al. (1983) proposed a two-part model that uses two equations to separate the modeling into two stages. The first stage refers to whether the response outcome is positive. Conditional on its being positive, the second stage refers to its level.

The first part is a binary model for the dichotomous event of having zero or positive values, such as the logistic regression model

$$\text{logit}[P(Y_i = 0)] = \boldsymbol{x}'_{1i}\boldsymbol{\beta}_1$$

Conditional on a positive value, the second part assumes a log-normal distribution; that is,

$$\log(y_i|y_i > 0) = \boldsymbol{x}'_{2i}\boldsymbol{\beta}_2 + \epsilon_i$$

where ϵ_i is distributed as $N(0, \sigma^2)$. The likelihood function for this

two-part model is

$$\begin{aligned}\ell(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma) &= \Big[\prod_{y_i=0} P(y_i = 0)\Big]\Big[\prod_{y_i>0} P(y_i > 0) f(y_i|y_i > 0)\Big] \\ &= \Big[\prod_{y_i=0} \frac{e^{\boldsymbol{x}'_{1i}\boldsymbol{\beta}_1}}{1 + e^{\boldsymbol{x}'_{1i}\boldsymbol{\beta}_1}}\Big] \\ &\quad \Big[\prod_{y_i>0} \frac{1}{1 + e^{\boldsymbol{x}'_{1i}\boldsymbol{\beta}_1}} \sigma^{-1}\phi\Big(\frac{\log(y_i) - \boldsymbol{x}'_{2i}\boldsymbol{\beta}_2}{\sigma}\Big)\Big]\end{aligned}$$

Duan et al. (1983) showed that the likelihood function has a unique global maximum. ML calculations are relatively simple, because the likelihood function factors into two terms. The first term has only the logit model parameters,

$$\ell_1(\boldsymbol{\beta}_1) = \Big[\prod_{y_i=0} e^{\boldsymbol{x}'_{1i}\boldsymbol{\beta}_1}\Big]\Big[\prod_{i=1}^{n} \frac{1}{1 + e^{\boldsymbol{x}'_{1i}\boldsymbol{\beta}_1}}\Big]$$

The second term involves only the parameters of the second model part,

$$\ell_2(\boldsymbol{\beta}_2, \sigma) = \prod_{y_i>0} \sigma^{-1}\phi\Big(\frac{\log(y_i) - \boldsymbol{x}'_{2i}\boldsymbol{\beta}_2}{\sigma}\Big)$$

One can obtain ML estimates by separately maximizing the two terms. Duan et al. (1983) applied this model to describe demand for medical care. For another application, see Grytten, Holst, and Laake (1993).

2.3 Sample selection models

Heckman (1974, 1979) extended the Tobit model to a two-part model. His model has been commonly applied to model sample selection and the related potential bias. There are many variants of sample selection models. We use the version by van de Ven and van Praag (1981) to illustrate. For observation i, let $\{(u_{1i}, u_{2i})\}$ be iid from a bivariate $N(\mathbf{0}, \boldsymbol{\Sigma})$ distribution, where

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}$$

The model assumes that

$$I_i = \boldsymbol{x}'_{1i}\boldsymbol{\beta}_1 + u_{1i},$$

$$y_i^* = \boldsymbol{x}_{2i}'\boldsymbol{\beta}_2 + u_{2i},$$

$$\begin{aligned} y_i &= \exp(y_i^*) \quad \text{if } I_i > 0, \\ &= 0 \quad \text{if } I_i \leq 0 \end{aligned}$$

When $I_i > 0$, $y_i > 0$ is observed and $y_i^* = \log(y_i)$; when $I_i \leq 0$, $y_i = 0$ is observed and y_i^* is 'missing'. The covariate and parameter vectors $(\boldsymbol{x}_{1i},\ \boldsymbol{\beta}_1)$ for I_i may differ from $(\boldsymbol{x}_{2i},\ \boldsymbol{\beta}_2)$ for y_i^*. Two estimation methods employed with this model are ML and a two-step procedure due to Heckman (1979).

For ML estimation, the likelihood function of the model is given by

$$\begin{aligned} \ell(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\Sigma}) &= \Big[\prod_{y_i=0} P(I_i \leq 0)\Big]\Big[\prod_{y_i>0} f(y_i^*|I_i > 0)P(I_i > 0)\Big] \\ &= \Big[\prod_{y_i=0} P(I_i \leq 0)\Big]\Big[\prod_{y_i>0} \int_0^\infty f(y_i^*, I_i)dI_i\Big] \\ &= \Big[\prod_{y_i=0} \{1 - \Phi\Big(\frac{\boldsymbol{x}_{1i}'\boldsymbol{\beta}_1}{\sigma_1}\Big)\}\Big] \\ &\quad \times\Big[\prod_{y_i>0} \Phi\{\Big(\frac{\boldsymbol{x}_{1i}'\boldsymbol{\beta}_1}{\sigma_1} + \frac{\log(y_i) - \boldsymbol{x}_{2i}'\boldsymbol{\beta}_2}{\sigma_{12}^{-1}\sigma_1\sigma_2^2}\Big) \\ &\quad \times(1 - \sigma_{12}^2\sigma_1^{-2}\sigma_2^{-2})^{-\frac{1}{2}}\}\sigma_2^{-1}\phi\Big(\frac{\log(y_i) - \boldsymbol{x}_{2i}'\boldsymbol{\beta}_2}{\sigma_2}\Big)\Big] \end{aligned}$$

An iterative method can be used to find the ML estimates.

Heckman's two-step procedure does not perform as well as the ML estimators. But this method is very simple and easy to implement. It is widely used and has become the standard estimation procedure for empirical microeconometrics studies. With the two-step procedure, the subsample regression function for Y_i^* is

$$E[Y_i^*|\boldsymbol{x}_{2i}, I_i > 0] = \boldsymbol{x}_{2i}'\boldsymbol{\beta}_2 + E[u_{2i}|u_{1i} > -\boldsymbol{x}_{1i}'\boldsymbol{\beta}_1] = \boldsymbol{x}_{2i}'\boldsymbol{\beta}_2 + \frac{\sigma_{12}}{\sigma_1}\lambda_i \tag{1}$$

where $\lambda_i = \phi(z_i)/\Phi(z_i)$, and $z_i = \boldsymbol{x}_{1i}'\boldsymbol{\beta}_1/\sigma_1$. So, we have

$$\begin{aligned} \log(Y_i) &= E[Y_i^*|\boldsymbol{x}_{2i}, I_i > 0] + \epsilon_i \\ &= \boldsymbol{x}_{2i}'\boldsymbol{\beta}_2 + \frac{\sigma_{12}}{\sigma_1}\lambda_i + \epsilon_i, \end{aligned}$$

where Heckman (1979) showed that ϵ_i has mean 0 and variance $\sigma_2^2[(1 - \rho^2) + \rho^2(1 + z_i\lambda_i - \lambda_i^2)]$, where $\rho^2 = \sigma_{12}^2/(\sigma_1^2\sigma_2^2)$. One can

estimate the parameters $\boldsymbol{\beta}_1$ and σ_1 by a probit model using the full sample. Therefore, z_i and hence λ_i can be easily estimated. The estimated value of λ_i is used as a regressor in equation (1). Then one can estimate $\boldsymbol{\beta}_2$ using least squares.

Duan et al. (1983, 1984) pointed out that the model has poor numerical and statistical properties. The likelihood function may have non-unique local maxima (Olsen 1975), and computations are more involved than in the Duan et al. (1983) two-part model. The model relies on untestable assumptions in that the censored data are unobservable, so standard diagnostic methods based on the empirical error distribution cannot be applied. When a high correlation exists between λ and $\boldsymbol{x}_2$, the estimator in the sample selection model is very nonrobust. Some researchers have suggested that $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ should not have variables in common, but this is not realistic in practice.

Both the Duan et al. (1983) two-part model and Heckman's sample selection model use two equations to separately model whether the outcome is positive and the magnitude of a positive response. The sample selection model posits an underlying bivariate normal error. It estimates an unconditional equation that describes the level that subjects would have if they all had outcomes. The two-part model estimates a conditional equation that describes only the level of outcomes for those that truly are positive. The econometrics literature contains discussion comparing the sample selection model and the two-part model. See, for instance, Duan et al. (1983, 1984), Manning et al. (1987), and Leung and Yu (1996).

2.4 Compound Poisson exponential dispersion models

Jørgensen (1987, 1997) proposed using a single distribution from the exponential dispersion family to analyze semicontinuous data. This distribution is a type of compound Poisson distribution. The exponential dispersion family, which is used in generalized linear models, has form

$$f(y_i;\, \theta_i, \phi) = c(y_i, \phi) \exp\left(\frac{\theta_i y_i - b(\theta_i)}{\phi}\right)$$

It is characterized by its variance function $v(\mu_i)$, expressed in terms of the mean μ_i (Jørgensen 1987). For this family, θ relates to μ by $\mu = \partial b(\theta)/\partial\theta$. An important class of exponential dispersion models uses the power function, $v(\mu) = \mu^p$. When $p = 1$, this is the Poisson distribution.

Jørgensen (1997) applied this family for $1 < p < 2$, for which

$$b(\theta_i) = (\frac{\alpha - 1}{\alpha})(\frac{\theta_i}{\alpha - 1})^{\alpha}$$

where $\alpha = (p-2)/(p-1)$, and

$$c(y_i, \phi) = \begin{cases} \frac{1}{y_i} \sum_{n=1}^{\infty} \frac{b^n(-\phi/y_i)}{\phi^n \Gamma(-\alpha n) n!} & y_i > 0 \\ 1 & y_i = 0 \end{cases}$$

For this distribution,

$$\mu_i = \partial b(\theta_i)/\partial \theta_i = \left(\frac{\theta_i}{\alpha - 1}\right)^{\alpha - 1}$$

Jørgensen (1997) showed that when $1 < p < 2$, this distribution results from the compound Poisson construction,

$$Y_i = \sum_{j=0}^{N_i} W_{ij}$$

where N_i has a Poisson$(b(\theta_i)/\phi)$ distribution and W_{ij} has a gamma $(\alpha\phi/\theta_i, -\alpha)$ distribution. When N_i and $\{W_{ij}\}$ are independent, $P(Y_i = 0) = P(N_i = 0)$. Given $N_i > 0$, the distribution of Y_i is continuous on the positive real line.

With link function g, one can specify a model for the mean response as $g(\mu_i) = \boldsymbol{x}_i'\boldsymbol{\beta}$. Obtaining the ML estimator for $\boldsymbol{\beta}$ does not involve $c(y_i, \phi)$. When p is known, this model can be fitted with software for generalized linear models. Normally, however, p would itself be unknown and need to be estimated. Since it occurs (through α) in the infinite sum and gamma function in $c(y_i, \phi)$, estimating it can be computationally difficult (Jørgensen 1987). Alternative moment-based estimation may perform well. Tweedie (1984) suggested an estimate of p based on a single random sample as $\hat{p} = \hat{k}_1 \hat{k}_3 \hat{k}_2^{-2}$, where $\hat{k}_t$ is an estimate of cumulant t of the distribution. Jørgensen proposed a possible generalization of this approach for a regression model. Let $\boldsymbol{y}$ and $\hat{\boldsymbol{\mu}}$ represent vectors of observations and fitted values. A moment estimator for ϕ is $\hat{\phi} = \boldsymbol{X}^2/(n-k)$, where k is the number of unknown parameters amd $\boldsymbol{X}^2 = (\boldsymbol{y} - \hat{\boldsymbol{\mu}})^T V(\hat{\boldsymbol{\mu}})^{-1}(\boldsymbol{y} - \hat{\boldsymbol{\mu}})$.

2.5 Ordinal threshold models

Saei, Ward, and McGilchrist (1996) suggested grouping the possible outcome values into k ordered categories and applying an ordinal response model. Let Y_g be the grouped response variable. The threshold model for an ordinal response posits an unobservable variable Z, such that one observes $Y_g = j$ (i.e., in category j) if Z is between θ_{j-1} and θ_j. Suppose that Z has a cumulative distribution function $G(z - \eta)$, where η is related to explanatory variables by

$$\eta = \boldsymbol{x}'\boldsymbol{\beta}$$

Then,

$$P(Y_g \leq j) = P(Z \leq \theta_j) = G(\theta_j - \boldsymbol{x}'\boldsymbol{\beta})$$

The threshold model then follows, by which

$$G^{-1}[P(Y_g \leq j; \boldsymbol{x})] = \theta_j - \boldsymbol{x}'\boldsymbol{\beta}, \quad j = 1, 2, \ldots, k-1$$

That is, the inverse of the *cdf* serves as the link function.

In application with semicontinuous data and a clump at 0, one would take the first category to be the 0 outcome, and then one would select cutpoints on the positive outcome scale to define the other $k-1$ categories. Assuming that G is logistic leads to a logit model for the cumulative probabilities, called a cumulative logit model. Assuming that G is normal leads to a cumulative probit model (McCullagh 1980). A score test is available to check the assumption that covariate effects are the same for each cutpoint (Peterson and Harrell 1990). Chang and Pocock (2000) applied the cumulative logit model for modeling the amount of personal care for the elderly.

This model has the simplicity of a single model to handle the clump at 0 and the positive outcomes. Elements of $\boldsymbol{\beta}$ summarize effects overall, rather than conditional on the response being positive. For instance, to compare different groups that are levels of the explanatory variables, one can use $\hat{\boldsymbol{\beta}}$ directly, whereas for two-part models one needs to average results from the two components of the model to make an unconditional comparison (e.g., to estimate $E(Y)$ for the groups). Two obvious concerns with this model are that the way the positive scale is collapsed into categories is arbitrary, and by grouping the data one loses some information.

2.6 Advantages and disadvantages of existing approaches

The Tobit model was the first to deal with semicontinuous data. The sample selection model extends the Tobit model to allow different coefficients to affect the two components. Both models assume an underlying normal random variable that is censored by a random mechanism. These models are sometimes suitable for modeling a limited or censored response variable. When zeros represent actual outcome values instead of censored or missing values, the underlying normal assumption becomes dubious. By contrast, the Duan et al. (1983) two-part model has several appealing properties, including a well-behaved likelihood function and more appropriate interpretations than the Tobit and Heckman models if the zeros are true values.

The compound Poisson exponential dispersion model makes it possible to analyze data with a single model that includes both aspects described in the two-part model. In this sense, it is relatively simple. Given the power p in the variance function, this model is easy to fit, but otherwise the model seems problematic. It does not seem to have received attention in practice other than in Jørgensen's work. Ordinal response models also can model the zero and non-zero values in one model, and they are simple to fit. A drawback is that they model grouped data instead of the original data.

Of these models, it seems to us that the Duan et al. (1983) two-part model is a reasonable choice for many applications. Compared with other models we've discussed, this model addresses the data in their original form, is simple to fit, and is relatively simple to interpret.

3 Models for Zero-Inflated Count Data

Count responses with a relatively large clump at zero can occur in many situations (e.g., Cameron and Trivedi 1998, pp. 10-15). Having a large number of observations at zero is not by itself sufficient to rule out a particular discrete distribution. However, often the remaining counts show considerable variability, which is inconsistent with the Poisson distribution (for which the mean determines both the variance and the probability at 0). This may be caused by overdispersion due to unobserved heterogeneity. Then, a distribution that allows the Poisson mean to vary at fixed values of predictors may be appropri-

ate. Examples are the negative binomial regression model (which can be derived with a gamma mixture of Poisson means) and the generalized linear mixed model that adds a normal random effect to a model for the log of the Poisson mean. See, for instance, Cameron and Trivedi (1998) and Chapter 13 of Agresti (2002) for discussion of such approaches.

Sometimes such simple models for overdispersion are themselves inadequate. For instance, the data might be bimodal, with a clump at zero and a separate hump around some considerably higher value. This might happen for variables for which a certain fraction of the population necessarily has a zero outcome, and the remaining fraction follows some distribution having positive probability of a zero outcome. This happens for variables referring to the number of times one takes part in a certain activity, when some subjects never do so and others may occasionally not do so. Examples are the number of papers one published in the previous year (for a sample of professors), and the number of times one exercised in a gym in the previous month. For such zero-clumped data, standard discrete distributions are suspect. The above representation of two types of subjects leads naturally to a mixture model, some examples of which are presented in this section on the modeling of zero-inflated count data.

3.1 Zero-inflated discrete distributions

Lambert (1992) introduced zero-inflated Poisson (ZIP) regression models to account for overdispersion in the form of excess zero counts for the Poisson distribution. Since her article, zero-inflated discrete models have been developed and applied in the econometrics and statistics literature.

Lambert's model treats the data as a mixture of zeros and outcomes of Poisson variates. For subject i, she assumed that

$$Y_i \sim \begin{cases} 0 & \text{with probability } p_i \\ \text{Poisson}(\lambda_i) & \text{with probability } 1-p_i \end{cases}$$

The resulting distribution has

$$\begin{aligned} P(Y_i = 0) &= p_i + (1-p_i)e^{-\lambda_i}, \\ P(Y_i = j) &= (1-p_i)\frac{e^{-\lambda_i}\lambda_i^j}{j!}, \qquad j = 1, 2, \ldots \end{aligned}$$

With explanatory variables, the parameters are themselves modeled by

$$\text{logit}(p_i) = \boldsymbol{x}'_{1i}\boldsymbol{\beta}_1 \quad \text{and} \quad \log(\lambda_i) = \boldsymbol{x}'_{2i}\boldsymbol{\beta}_2$$

The log likelihood function is

$$\begin{aligned} L(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) &= \sum_{y_i=0} \log[e^{\boldsymbol{x}'_{1i}\boldsymbol{\beta}_1} + \exp(-e^{\boldsymbol{x}'_{2i}\boldsymbol{\beta}_2})] \\ &\quad + \sum_{y_i>0} (y_i \boldsymbol{x}'_{2i}\boldsymbol{\beta}_2 - e^{\boldsymbol{x}'_{2i}\boldsymbol{\beta}_2}) \\ &\quad - \sum_{i=1}^{n} \log(1 + e^{\boldsymbol{x}'_{1i}\boldsymbol{\beta}_1}) - \sum_{y_i>0} \log(y_i!) \end{aligned}$$

A latent class construction that yields this model posits an unobserved binary variable Z_i. When $Z_i = 1$, $y_i = 0$, and when $Z_i = 0$, Y_i is Poisson(λ_i). Lambert (1992) suggested using the EM algorithm for ML estimation of the parameters, treating z_i as a missing value.

Hall (2000) adapted Lambert's method to an upper-bounded count setting to yield a zero-inflated binomial model. With upper bound for Y_i of n_i, he took

$$Y_i \sim \begin{cases} 0 & \text{with probability } p_i \\ \text{binomial}(n_i, \pi_i) & \text{with probability } 1 - p_i \end{cases}$$

He modeled p_i with $\text{logit}(p_i) = \boldsymbol{x}'_{1i}\boldsymbol{\beta}_1$ and modeled π_i with $\text{logit}(\pi_i) = \boldsymbol{x}'_{2i}\boldsymbol{\beta}_2$, using the EM algorithm to obtain ML estimates.

In practice, overdispersion is common with count data, even conditional on a positive count or for a component of a latent class model. The equality of mean and variance assumed by the ZIP model, conditional on $Z_i = 0$, is often not realistic. Zero-inflated negative binomial models would likely often be more appropriate than ZIP models. Grogger and Carson (1991) used zero-truncated Poisson models to fit data simulated from zero-truncated negative binomial distributions. They observed biases of estimated parameters up to 30 percent. Similar arguments extend to zero-inflated models. With an inappropriate Poisson assumption, standard error estimates can be biased very dramatically. Ridout, Hinde and Demetrio (2001) provided a score test for testing zero-inflated Poisson models against the zero-inflated negative binomial alternative. For an application of the zero-inflated negative binomial model, see Shankar, Milton, and Mannering (1997).

With more than a single unusually high probability, extensions of zero-inflated count models may be needed. For instance, in studying

Swedish female fertility, Melkersson and Rooth (2000) inspected the number of births for a sample of women. They found more 0 and 2 outcomes than expected in a standard count data model. They used a multinomial logit model to estimate the extra probabilities of zero and two children.

3.2 Hurdle models

The hurdle model is a two-part model for count data proposed by Mullahy (1986). One part of the model is a binary model, such as logistic or probit regression, for whether the response outcome is zero or positive. If the outcome is positive, the "hurdle is crossed." Conditioning on a positive outcome, to analyze its level the second part uses a truncated model that modifies an ordinary distribution by conditioning on a positive outcome. This might be a truncated Poisson or truncated negative binomial. Applications of such models have been given by Pohlmeier and Ulrich (1995), Arulampalam and Booth (1997), and Gurmu and Trivedi (1996).

Suppose we use a logistic regression for the binary process and a truncated Poisson model for the positive outcome; that is,

$$\text{logit}[P(Y_i = 0)] = \boldsymbol{x}'_{1i}\boldsymbol{\beta}_1 \quad \text{and} \quad \log(\lambda_i) = \boldsymbol{x}'_{2i}\boldsymbol{\beta}_2$$

The log likelihood then has two components:

$$\begin{aligned} L_1(\boldsymbol{\beta}) &= \sum_{y_i=0}[\log P_1(y_i = 0; \boldsymbol{\beta}_1, \boldsymbol{x}_{1i})] \\ &\quad + \sum_{y_i>0}[\log(1 - P_1(y_i = 0; \boldsymbol{\beta}_1, \boldsymbol{x}_{1i}))] \\ &= \sum_{y_i=0} \boldsymbol{x}'_{1i}\boldsymbol{\beta}_1 - \sum_{i=1}^{n} \log(1 + e^{\boldsymbol{x}'_{1i}\boldsymbol{\beta}_1}) \end{aligned}$$

is the log-likelihood function for the binary process, and

$$L_2(\boldsymbol{\beta}_2) = \sum_{y_i>0}[y_i \boldsymbol{x}'_{2i}\boldsymbol{\beta}_2 - e^{\boldsymbol{x}'_{2i}\boldsymbol{\beta}_2} - log(1 - e^{-e^{\boldsymbol{x}'_{2i}\boldsymbol{\beta}_2}})] - \sum_{y_i>0} \log(y_i!)$$

is the log-likelihood function for the truncated model. The joint log-likelihood function is

$$L(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = L_1(\boldsymbol{\beta}_1) + L_2(\boldsymbol{\beta}_2)$$

One can maximize this by separately maximizing L_1 and L_2.

In some applications, the data may have a long right tail reflecting some extremely large positive counts. Gurmu (1997) proposed a semi-parametric hurdle model for a highly skewed distribution of counts. It is based on a Laguerre series expansion for the unknown density of the unobserved heterogeneity.

3.3 Finite mixture models

Another approach for zero-inflated count data uses a finite mixture model. It assumes that the response comes from a mixture of several latent distributions. With q latent groups, the mixture density is

$$f(y_i;\boldsymbol{\theta}) = \sum_{j=1}^{q} \pi_j f_j(y_i;\theta_j), \quad y = 0, 1, 2, \ldots$$

where π_j is the true proportion in group j, $f_j(y_i;\theta_j)$ is the mass function (e.g., Poisson or negative binomial) for group j, and $\{\pi_j\}$ and $\{\theta_j\}$ are unknown parameters. The zero-inflated count models of Sec. 3.1 are special cases of the finite mixture model in which one of the mixture mass functions is degenerate at zero. The more general mixture model allows for additional population heterogeneity but avoids the sharp dichotomy between the population of zeros and non-zero counts.

One approach to fitting a finite mixture model relates it to latent class analysis (Aitkin and Rubin 1985). Let d_{ij} denote an indicator to represent whether y_i comes from latent group j, with $\sum_j d_{ij} = 1$. Assume that $\{(y_i, d_{i1}, \ldots, d_{iq}),\ i = 1, \ldots, n\}$ are independent, such that $\{d_{ij}, j = 1, \ldots, q\}$ have the multinomial distribution

$$\prod_{j=1}^{q} \pi_j^{d_{ij}}$$

and conditional on their values, y_i has probability mass function

$$\sum_{j=1}^{q} d_{ij} f(y_i;\theta_j) = \prod_{j=1}^{q} f(y_i;\theta_j)^{d_{ij}}, \quad y_i = 0, 1, 2, \ldots$$

Then, the likelihood function is

$$\ell(\boldsymbol{\theta}, \boldsymbol{\pi}) = \prod_{i=1}^{n} \left[\sum_{j=1}^{q} \pi_j^{d_{ij}} f(y_i;\theta_j)^{d_{ij}} \right]$$

Treating $\{d_{ij}\}$ as missing data, one can use the EM algorithm to fit the model.

Deb and Trivedi (1997) used a finite mixture model to study the demand for medical care by the elderly. They found that a two-point mixture negative binomial model fits better than the standard negative binomial model and its hurdle extension. Wedel et al. (1993) applied this method to analyze the effects of direct marketing on book selling. Gerdtham and Trivedi (2001) used it in studying the equity issue in Swedish health care.

3.4 Neyman type A distribution

Dobbie and Welsh (2001) proposed modeling zero-inflated count data using the Neyman type A distribution. This distribution is a compound Poisson-Poisson mixture. For observation i, let N_i denote a Poisson variate with expected value λ_i. Conditional on N_i, let W_{it} $(t = 1, \ldots, N_i)$ denote independent observations from a Poisson distribution with expected value ϕ_i. The model expresses Y_i using the decomposition,

$$Y_i = \sum_{t=0}^{N_i} W_{it} \quad , \quad i = 1, 2, \ldots, n$$

The probability mass function for Y_i is

$$\begin{aligned} P(Y_i = y_i) &= \sum_{j=0}^{\infty} \left[P(\sum_{t=0}^{N_i} W_{it} = y_i | N_i = j) P(N_i = j) \right] \\ &= \sum_{j=0}^{\infty} \left[\frac{e^{-j\phi_i}(j\phi_i)^{y_i}}{y_i!} \right] \left[\frac{e^{-\lambda_i}\lambda_i^j}{j!} \right] \\ &= \frac{e^{-\lambda_i}\phi_i^{y_i}}{y_i!} \sum_{j=0}^{\infty} \frac{(\lambda_i e^{-\phi_i})^j j^{y_i}}{j!} \end{aligned}$$

Using this distribution, one can form a model that relates λ_i and ϕ_i to explanatory variables through

$$\log(\lambda_i) = \boldsymbol{x}'_{1i}\boldsymbol{\beta}_1,$$

$$\log(\phi_i) = \boldsymbol{x}'_{2i}\boldsymbol{\beta}_2$$

Since $E(Y_i) = \lambda_i \phi_i$,

$$\log[E(Y_i)] = \log(\lambda_i) + \log(\phi_i) = \boldsymbol{x}'_{1i}\boldsymbol{\beta}_1 + \boldsymbol{x}'_{2i}\boldsymbol{\beta}_2$$

Dobbie and Welsh used a four-step procedure with the Newton-Raphson algorithm to estimate parameters, iterating between estimating $\boldsymbol{\beta}_1$ for a given $\boldsymbol{\beta}_2$ and estimating $\boldsymbol{\beta}_2$ for a given $\boldsymbol{\beta}_1$. The infinite sums in the density function make model-fitting complicated. They applied it to model the abundance of Leadeater's Possum in mountain ash forests of southeastern Australia. Here, λ_i denotes the mean number of possum clusters per site, and ϕ_i denotes the average number of possums per cluster.

3.5 Advantages and disadvantages of existing approaches

The zero-inflated model and the hurdle model are similar. The zero-inflated models are more natural when it is reasonable to think of the population as a mixture, with one set of subjects that necessarily has a 0 response. However, they are more complex to fit, as the model components must be fitted simultaneously. By contrast, one can separately fit the two components in the hurdle model. The hurdle model is also suitable for modeling data with *fewer* zeros than would be expected under standard distributional assumptions.

The finite mixture model is semi-parametric. If the observations can realistically be viewed as being drawn from different populations, this approach is attractive. A potential disadvantage with this model is that it may overestimate the number of components when there is a lack of model fit. The Neyman type A model makes it possible to fit the data using a single distribution. However, it is not a member of the exponential family, so the mathematical and inferential advantages associated with this family are not available, and model fitting is complicated by the infinite sum in the mass function.

4 Modeling Repeated Measurement of Zero-Clumped Data

Compared with the substantial literature on cross-sectional observations of data with clumping at zero, few papers have discussed the

modeling of clustered, correlated observations, such as occur with longitudinal data. This section surveys this literature.

4.1 Repeated measurement of semicontinuous data

Cowles, Carlin, and Connett (1996) and Hajivassiliou (1994) extended the Tobit model and the sample selection model to longitudinal data. Both models assume an underlying normal distribution, which is dubious in most applications, especially when zeros represent actual responses instead of censored or missing values. We do not discuss their approaches here. Olsen and Schafer (2001) extended the two-part model of Duan et al. (1983) to longitudinal data. We describe their model next.

Let y_{ij} be the semicontinuous response for subject (or cluster) i ($i = 1, \ldots, n$) at occasion j ($j = 1, \ldots, t_i$). The first part of the model is a logistic random effects model for the dichotomous event of having zero or positive values. Suppose that

$$y_{ij} \begin{cases} = 0 & \text{with probability } p_{ij} \\ \neq 0 & \text{with probability } 1 - p_{ij} \end{cases}$$

Let $\{\boldsymbol{c}_i\}$ be random effects to account for within-subject correlation. Conditional on $\boldsymbol{c}_i$, we assume that

$$\text{logit}(p_{ij}) = \boldsymbol{x}'_{1ij}\boldsymbol{\beta}_1 + \boldsymbol{z}'_{1ij}\boldsymbol{c}_i$$

where $\boldsymbol{x}_{1ij}$ and $\boldsymbol{z}_{1ij}$ are covariate vectors pertaining to the fixed effects $\boldsymbol{\beta}$ and random effects $\boldsymbol{c}_i$. In practice, the simple random intercept form of model is often adequate, in which $\boldsymbol{c}_i = c_i$ is univariate and $\boldsymbol{z}_{1ij} = 1$.

In the second part of the model, let

$$V_{ij} = \begin{cases} y_{ij}, & \text{if } y_{ij} > 0 \\ \text{unspecified}, & \text{if } y_{ij} = 0 \end{cases}$$

When V_{ij} is positive, conditional on a random effect $\boldsymbol{d}_i$ the model assumes that V_{ij} follows a log-normal distribution. Thus, the model for the positive outcomes is

$$\log(V_{ij}) = \boldsymbol{x}'_{2ij}\boldsymbol{\beta}_2 + \boldsymbol{z}'_{2ij}\boldsymbol{d}_i + \epsilon_{ij}$$

where the residuals $\{\epsilon_{ij}\}$ are assumed to be independent from $N(0, \sigma^2)$. Again, often the simple random intercept form of model is often adequate, in which $\boldsymbol{d}_i = d_i$ is univariate and $\boldsymbol{z}_{2ij} = 1$.

When the response is observed at repeated times, a high level of a positive response at one time may affect the probability of a positive outcome at another time. So, one can tie the two parts of the model together by taking the random effects from the two parts as jointly normal and correlated,

$$\boldsymbol{b}_i = \begin{pmatrix} \boldsymbol{c}_i \\ \boldsymbol{d}_i \end{pmatrix} \sim N\left(\mathbf{0},\ \boldsymbol{\Sigma}\right)$$

where

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{cc} & \boldsymbol{\Sigma}_{cd} \\ \boldsymbol{\Sigma}_{dc} & \boldsymbol{\Sigma}_{dd} \end{pmatrix}$$

To fit the model, one first obtains a marginal likelihood by integrating out the random effects. However, these integrals are analytically intractable, so the marginal likelihood does not have a closed-form expression. Numerical or stochastic approximation of the integrals is needed, as in the fitting of generalized linear mixed models (e.g., Agresti 2002, Chapter 12). With univariate random intercepts, numerical approximation using Gauss-Hermite quadrature, which approximates the integral by a finite sum, should be adequate. Then one can maximize the approximated likelihood using standard optimization methods such as Newton–Raphson.

Olsen and Schafer (2001) studied many fitting methods. They compared Markov chain Monte Carlo (MCMC), the EM algorithm, penalized quasi-likelihood (PQL), Gauss-Hermite quadrature, and Laplace approximations. Simulations by Raudenbush et al. (2000) showed that a high-order Laplace approximation can be as accurate as Gauss-Hermite quadrature yet is much faster than the other methods. Olsen and Schafer noted that it took MCMC and EM algorithms more than one day to obtain accurate estimates for their example, while the sixth-order Laplace method needed less than one minute.

Saei et al. (1996) extended the ordinal threshold model to analyze clustered semicontinuous data. Again, this requires breaking the continuous scale into categories. Let $y_{ij,g}$ be the grouped response for observation j on subject i. Let G be the cumulative distribution function for an underlying unobservable variable. For outcome category k, the model assumes

$$P(Y_{ij,g} \leq k) = G(\theta_k - \eta_{ij})$$

where

$$\eta_{ij} = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\boldsymbol{b}_i$$

for a vector $\boldsymbol{b}_i \sim N(\boldsymbol{0}, \boldsymbol{\Sigma})$ of random effects that account for within-subject correlation. They took G to be the standard normal *cdf*, yielding a cumulative probit model, and they used penalized quasi-likelihood (PQL) to fit the model.

4.2 Repeated measurement of zero-inflated data

As with semicontinuous data, there is little literature on modeling clustered zero-inflated count data. Hall (2000) extended the zero-inflated Poisson and zero-inflated binomial models to handle longitudinal data, adding random effects to account for the within-subject dependence.

Hall assumed that $Y_{ij} = 0$ with probability p_{ij} and $Y_{ij} \sim \text{Poisson}(\lambda_{ij})$ with probability $1 - p_{ij}$. The parameters p_{ij} and λ_{ij} are modeled by

$$\text{logit}(p_{ij}) = \boldsymbol{x}'_{1ij}\boldsymbol{\beta}_1,$$
$$\log(\lambda_{ij}) = \boldsymbol{x}'_{2ij}\boldsymbol{\beta}_2 + b_i$$

where $b_i \sim N(0, \sigma^2)$ is a random effect. The Poisson(λ_{ij}) distribution applies conditional on b_i; unconditionally, there is overdispersion relative to the Poisson when $\sigma > 0$. The log-likelihood function for the longitudinal zero-inflated Poisson model is

$$L(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma) = \sum_{i=1}^{n} \left[\log \int_{-\infty}^{+\infty} [\prod_{j=1}^{t_i} P(Y_{ij} = y_{ij}|b_i)]\phi(b_i)db_i \right]$$

Hall employed the EM algorithm with Gauss-Hermite quadrature to fit the model. In a corresponding longitudinal zero-inflated binomial model, Hall assumed that Y_{ij} is binomial(n_i, π_{ij}) with probability $1 - p_{ij}$ (conditional on a random effect b_i), where $\text{logit}(\pi_{ij}) = \boldsymbol{x}'_{2ij}\boldsymbol{\beta}_2 + b_i$.

Note that Hall's model does not have a random effect for the part of the model determining the zero inflation. By contrast, Yau and Lee (2001) proposed adding a pair of uncorrelated normal random effects (b_i, c_i) for the two components of a hurdle model. They used a logistic model for the probability p_{ij} of a positive outcome. Conditional on positive outcome, they applied a loglinear model for the mean λ_{ij} in a truncated (conditionally) Poisson distribution. That is,

$$\text{logit}(p_{ij}) = \boldsymbol{x}'_{1ij}\boldsymbol{\beta}_1 + b_i,$$

$$\log(\lambda_{ij}) = \boldsymbol{x}'_{2ij}\boldsymbol{\beta}_2 + c_i$$

With uncorrelated random effects, the two components of the hurdle model can be fitted separately. Yau and Lee used a penalized quasi-likelihood approach for this.

5 An Application with Two Boundary Clumps

The form of data discussed in this article usually has only a single clump, occurring at zero, or a clump at zero and a mound around some larger value. Some applications, however, have data with two clumps, one at zero and one at the maximum possible response value.

An application in which such data commonly occur is in the study of patient compliance. Compliance is usually defined as the extent to which a subject's behavior (in terms of taking medications, following diets, or executing lifestyle changes) coincides with medical or health advice (Haynes, Taylor, and Sackett 1979). The response distribution usually has a proportion of subjects with 0% compliance, a proportion of subjects with 100% compliance, and other subjects having compliances spread between 0% and 100%. An appropriate model permits two clumps with positive probability at the extremes and treats the remaining scale as continuous.

Thus far, little attention has been paid to specialized models for compliance data. One possibility uses the logit to transform the response values between zero and one to the real line, as is often done with compositional data (Aitchison and Shen 1980) and continuous proportion data (Bartlett 1937). However, this transformation cannot handle 0% or 100% compliance. Another possibility is to use a quasi-likelihood approach with variance function $v(\mu) = [\mu(1-\mu)]^2$ (Wedderburn 1974), which corresponds to constant asymptotic variance for the logit of compliance. Another possibility is to use the Saei et al. (1996) approach with ordered categories for the outcomes, in which the extreme outcomes refer to the clumped outcomes. We are currently conducting research on models for this form of data.

6 Future Research

For cross-sectional nonnegative data with clumping at zero, we have surveyed a considerable amount of research. However, methods for

longitudinal data are less developed and could use more work. For instance, it may be of interest to extend the exponential dispersion model with $V(\mu) = \mu^p$ $(1 < p < 2)$ to longitudinal data analysis. So far this method has essentially been ignored even for cross-sectional analysis.

Or, one could develop further the ordinal threshold model for longitudinal data analysis. Saei et al. (1996) proposed a cumulative probit model to fit clustered semicontinuous data, and they used a penalized quasi-likelihood (PQL) approach to estimate the parameters. However, Breslow and Lin (1995) showed that the PQL estimates are biased and inconsistent for binomial responses when the random effects have large variance and the binomial denominator is small. We suspect that using the PQL approach to fit the cumulative probit model will have similar problems. The use of different distribution functions and more accurate estimation algorithms is open for future research. This type of model is also suitable for zero-inflated count data analysis. For ordinal threshold models, study is also needed about how close the parameters fitted by the grouped data tend to be to the ones fitted by ungrouped data.

In analyzing longitudinal count data with a zero-inflated Poisson model, Hall (2000) added a random intercept only to one component of the model. Yau and Lee (2001) added a pair of random effects to both components of a hurdle model. However, they assumed uncorrelated random effects and used PQL for model fitting. When the response is observed at several occasions, a high positive outcome at one time may affect the probability of a positive outcome at another time. These two processes are likely correlated and influenced by covariates in different ways. It makes sense to allow correlated random effects in the model, which then requires a more complex fitting process, preferably using ML.

For semicontinuous data analysis, most of the methods we mentioned assume that the positive continuous responses have a log-normal distribution. This need not be realistic, especially in applications in which some especially large observations create a right tail that is too long for a log-normal distribution. For instance, in a survey of medical care expenses, the right tail may be poorly modeled by the log normal. Semi-parametric methods may be appropriate for such highly skewed data. Finally, the modeling of compliance data and other types of data with more than one clump remains a fertile area for future research.

Acknowledgments

This research was partially supported by grants from NIH and NSF. The authors thank Dr. Alan Hutson for suggesting compliance data as a form of data possibly handled by some zero-clumped models.

References

Agresti, A. (2002), Categorical Data Analysis. 2nd Edition, Wiley.

Aitchison, J. and Shen, S. M. (1980), Logistic-normal distributions: Some properties and uses. Biometrika, **67**, 261–272.

Aitkin, M. and Rubin, D. B. (1985), Estimation and hypothesis testing in finite mixture models. Journal of the Royal Statistical Society, Series B, Methodological, **47**, 67–75.

Amemiya, T. (1973), Regression analysis when the dependent variable is truncated normal. Econometrica, **41**, 997–1016.

Amemiya, T. (1984), Tobit models: A survey. Journal of Econometrics, **24**, 3–61.

Arulampalam, W. and Booth, A.L. (1997), Who gets over the training hurdle? A study of the training experiences of young men and women in Britain. Journal of Population Economics, **10**, 197–217.

Bartlett, M. S. (1937), Some examples of statistical methods of research in agriculture and applied biology. Supplement to the Journal of the Royal Statistical Society, **4**, 137–183.

Breslow, N. E. and Lin, X. (1995), Bias correction in generalised linear mixed models with a single component of dispersion. Biometrika, **82**, 81–91.

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

Chang, B. -H., Pocock, S. (2000), Analyzing data with clumping at zero – an example demonstration. Journal of Clinical Epidemiology, **53**, 1036–1043.

Cowles, M. K. and Carlin, B. P. and Connett, J. E. (1996), Bayesian Tobit modeling of longitudinal ordinal clinical trial compliance data with nonignorable missingness. Journal of the American Statistical Association, **91**, 86–98.

Cragg, J. G. (1971), Some statistical models for limited dependent variables with application to the demand for durable goods. Econometrica, **39**, 829–844.

Deb, P. and Trivedi, P. K. (1997), Demand for medical care by the elderly: A finite mixture approach. Journal of Applied Econometrics, **12**, 313–336.

Dobbie, M. and Welsh, AH. (2001), Models for zero-inflated count data using the Neyman type a distribution. Statistical Modelling, **1**, 65–80.

Duan, N., Manning, W. G. Jr., Morris, C. N., and Newhouse, J. P. (1983), A comparison of alternative models for the demand for medical care (Corr: V2 P413). Journal of Business and Economic Statistics, **1**, 115–126.

Duan, N., Manning, W. G. Jr., Morris, C. N., and Newhouse, J. P. (1984), Choosing between the sample-selection model and the multi-part model. Journal of Business and Economic Statistics, **2**, 283–289.

Gerdtham, U. G. and Trivedi, P. K. (2001), Equity in swedish health care reconsidered: New results based on the finite mixture model. Health Economics, **10**, 565–572.

Green, P. J. (1984), Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives (with discussion). Journal of the Royal Statistical Society, Series B, Methodological, **46**, 149–192.

Grogger, J. T. and Carson, R. T. (1991), Models for truncated counts. Journal of Applied Econometrics, **6**, 225–238.

Gronau, R. (1974), Wage comparisons – a selectivity bias. Journal of Political Economy, **82**, 1119–1144.

Grytten, J., Holst, D., and Laake, P. (1993), Accessibility of dental services according to family income in A non-insured population. Social Science & Medicine, **37**, 1501–1508.

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

Gurmu, S. (1997), Semi-parametric estimation of hurdle regression models with an application to medicaid utilization. Journal of Applied Econometrics, **12**, 225–242.

Hajivassiliou, V. A. (1994), A simulation estimation analysis of the external debt crises of developing countries. Journal of Applied Econometrics, **9**, 109–131.

Hall, D. B. (2000), Zero-inflated poisson and binomial regression with random effects: a case study. Biometrics, **56**, 1030–1039.

Haynes, R. B., Taylor, D. W., and Sackett, D. L. (1979), Compliance in Health Care. Johns Hopkins University Press.

Heckman, J. (1974), Shadow prices, market wages, and labor supply. Econometrica,**42**, 679–694.

Heckman, J. (1979), Sample selection bias as a specification error. Econometrica, **47**, 153–161.

Jørgensen, B. (1987), Exponential dispersion models. Journal of the Royal Statistical Society, Series B, Methodological, **49**, 127–145.

Jørgensen, B. (1997), The theory of dispersion models. Chapman & Hall, Page 256.

Lambert, D. (1992), Zero-inflated poisson regression, with an application to defects in manufacturing. Technometrics, **34**, 1–14.

Leung, S. F. and Yu, S. (1996), On the choice between sample selection and two-part models. Journal of Econometrics, **72**, 197–229.

Manning, W. G. and Duan, N. and Rogers, W. H. (1987), Monte carlo evidence on the choice between sample selection and two-part models. Journal of Econometrics, **35**, 59–82.

McCullagh, P. (1980), Regression models for ordinal data. Journal of the Royal Statistical Society, Series B, Methodological, **42**, 109–142.

Melkersson, M. and Rooth, D. (2000), Modeling female fertility using inflated count data models. Journal of Population Economics, **13**, 189–203.

Mullahy, J. (1986), Specification and testing of some modified count data models. Journal of Econometrics, **33**, 341–365.

Olsen, MK. and Schafer, JL. (2001), A two-part random-effects model for semicontinuous longitudinal data. Journal of the American Statistical Association, **96**, 730–745.

Olsen, R. (1975), The analysis of two-variable models when one of the variables is dichotomous. Yale Unversity, Economics Dept. unpublished manuscript.

Peterson, B. and Harrell, F. E. (1990), Partial proportional odds models for ordinal response variables. Applied Statistics, **39**, 205–217.

Pohlmeier, W. and Ulrich, V. (1995), An econometric model of the two-part decision making process in the demand of health care. Journal of Human Resources, **30**, 339–361.

Powell, J. L. (1986), Symmetrically trimmed least squares estimation for tobit models. Econometrica, **54**, 1435–1460.

Raudenbush, S. W., Yang, M. -L. and Yosef, M. (2000), Maximum likelihood for generalized linear models with nested random effects via high-order laplace approximation. Journal of Computational and Graphical Statistics, **9**, 141–157.

Ridout, M. and Hinde, J. and Demetrio, CGB. (2001), A score test for testing a zero-inflated poisson regression model against zero-inflated negative binomial alternatives. Biometrics, **57**, 219–223.

Robinson, P. M. (1982), On the asymptotic properties of estimators of models containing limited dependent variables. Econometrica, **50**, 27–42.

Saei, A., Ward, J. and McGilchrist, C. A. (1996), Threshold models in a methadone programme evaluation. Statistics in Medicine, **15**, 2253–2260.

Shankar, V., Milton, J. and Mannering, F. (1997), Modeling accident frequencies as zero-altered probability processes: an empirical inquiry. Accident Analysis and Prevention, **29**, 829–837.

Tobin, J. (1958), Estimation of relationships for limited dependent variables. Econometrica, **26**, 24–36.

Tweedie, M. C. K. (1984), An index which distinguishes between some important exponential families. Statistics Applications and New Directions, Proceedings of the Indian Statistical Institute Golden Jubilee International Conference, Indian Statistical Institute (Calcutta), 579–604.

Van de Ven, W. and Van Praag, B. (1981), The demand for deductibles in private health insurance: a probit model with sample selection. Journal of Econometrics, **17**, 229–252.

Wedderburn, R. W. M. (1974), Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. Biometrika, **61**, 439–447.

Wedel, M., DeSarbo, W. S., Bult, J. R., and Ramaswamy, V. (1993), A latent class poisson regression model for heterogeneous count data. Journal of Applied Econometrics, **8**, 397–411.

Yau, K. K. and Lee, A. H. (2001), Zero-inflated poisson regression with random effects to evaluate an occupational injury prevention programme. Statistics in Medicine, **20**, 2907–2920.

Yoo, S. H., Kim, T. Y., and Lee, J. K. (2001), modeling zero response data from willingness to pay surveys – A semi-parametric estimation. Economics Letters, **71**, 191–196.

Zorn, C. J. W. (1998), An analytic and empirical examination of zero-inflated and hurdle poisson specifications. Sociological Methods & Research, **26**, 368–400.