Paper 250-2009

A Flexible Count Data Regression Model Using SAS® PROC NLMIXED

Nan-Ting Chou, University of Louisville, Louisville, KY
David Steenhard, LexisNexis, Dayton, OH

ABSTRACT

Count data regression models are used when the dependent variables are non-negative integers. The standard count data models are limited in their ability in handling the data distributions. Poisson and negative binomial distributions are commonly used in count models. Poisson distributions assumes equi-dispersed data (variance equals to the mean); and negative binomial regression models over-dispersed data (variance greater than the mean). While there are studies (Liu and Cela 2008; Tin 2008) provide zero-inflated and hurdle count data models in SAS, no study has provided a SAS program that allows for a comprehensive list of data distributions and modeling strategies. This paper presents a SAS® macro program that allows for a wide variety of count data distributions which can be used to model both under- and over-dispersed data. In addition, our SAS® macro program can handle data that has excess zeros (zero-inflated) in the sample. This SAS® macro is flexible in allowing one to estimate a variety of count regression models including: zero-inflated, hurdle, censored, truncated, finite mixture, semi-parametric, squared polynomial expansion, and generalized heterogeneous. We demonstrate this SAS® macro procedure by applying it to the number of takeover bids received by targeted firms. We also evaluate count models performance using goodness-of-fit test, Vuong's test, and information criteria test.

1. INTRODUCTION

Count data regression models are used when the dependent variable takes on non-negative integer values. Cameron and Trivedi (1996) and Long (1997) provide good overviews of count regression models. Count data models are widely used in empirical studies. Some recent research used count models are as follows. Yang (2007) uses a Poisson distribution count model to explore factors affecting the potential entry into an industry. Hellström *and* Nordström (2008) using the count data modeling to analyze household's choice of total number of nights to spend on monthly recreational trips. Nelson and Young (2008) studies the effects of various factors on alcohol advertising in magazines using the Poisson and negative binomial count regressions. *Czado et al. (2007) proposed* an extension of zero-inflated generalized Poisson regression models for count data. Guikema and Goffelt (2008) presents a count model based on Conway-Maxwell Poisson (COM) distribution that is useful for both underdispersed and overdispersed count data. Our paper presents a count regression model written in SAS macro that is capable of handling various types of count data distribution.

Applying linear regression to count data leads to inconsistent standard errors and may produce negative predictions for the dependent variable. Even with a logged dependent variable, the least squared estimates have these problems and are biased and inconsistent (King, 1989). Therefore count dependent variables require different modeling. The most common assumption of count data distribution is the Poisson distribution which restricts the data distribution to be equal-dispersion (the conditional variance equals the conditional mean). This stringent restriction cannot handle many empirical applications. Other modeling distributions have been developed. Mixed-Poisson distributions and negative binomial distributions have been widely used in situations where counts display overdispersion (conditional variance exceeds the conditional mean). For underdispersion (conditional variance is less than conditional mean) there are fewer modeling options. Since there is no model that handles only the underdispersed data, with underdispersed data we need to consider models that are flexible enough to cover both over- and under-dispersed data. Models that provide this flexibility include: the generalized event count (GEC(k)) model (Winkelmann and Zimmermann,1991 and 1995), double Poisson (Efron 1986), Poisson polynomial expansion, hurdle models (Mullahy 1986), and the generalized Poisson models (Famoye 1993, Famoye and Singh 2003) .

Another common problem with count model is that the number of zeros in a sample often exceeds the number predicted by the regression model. This happens because zeros may arise from two conditions. Zeros could represent things that would never occur or the events that would occur but did not during the time the data was collected. Zero-modified count models, namely, the zero-inflated model and hurdle

models are useful in addressing this issue and they can handle over- and under- dispersed data. We discuss the zero-modified count models in section 3 of this paper.

Many count variables are often censored or truncated. Zero truncated samples occur when observations enter the sample only after the first count occurs. While truncation can occur at any value, truncation at zero occurs most often in practice. Models allowing for censoring are required if observations (y_i ,$\boldsymbol{x}_i$) are available only for a restricted range of y_i, while those for $\boldsymbol{x}_i$ are always observable for all range of y_i. Hence, censored data involves loss of information less serious than truncated data. For censored and truncated count date, the model's log likelihood function needs to be specified accordingly.

To resolve the above problems with count data modeling, we developed a flexible count model using SAS® macro called *%countreg*. This macro allows one to select from a variety of count data distributions and techniques that treat overdispersion, underdispersion, zero-inflated, censored, and truncated count data. Although there are several procedures within the SAS software can be used to estimate count models, most of them are limited in some ways. The GENMOD, GLIMMIX and COUNTREG procedures are limited to the Poisson and the negative binomial distributions. The NLMIXED provides greater flexibility by specifying the log likelihood function of discrete count distributions. Using the NLMIXED procedure, Liu and Cela (2008) provided the hurdle, zero-inflated, and zero-inflated (tau) count models in addition to the Poisson and negative binomial regressions. They evaluate these five count models though an example of healthcare utilization.

In addition to specifying the five count data distributions presented in Liu and Cela (2008), our SAS® macro program specifies many other count data distributions such as: double Poisson, generalized Poisson, generalized Poisson constrained finite mixture, Poisson-Normal mixture, Poisson-inverse Gaussian mixture (Folks and Chhikara 1978), Poisson semi-parametric (Gurmu et. al. 1998), Poisson polynomial expansion, geometric Poisson (Polya-Aeeppli), negative binomial finite mixture, negative binomial constrained finite mixture, negative binomial polynomial expansion, Neyman Type A (Neyman,1939), and the generalized event count (GEC(k)) model. We specify these data distributions through the log likelihood option in the Model statement in PROC NLMIXED procedure in a SAS® macro. This SAS® macro program provides count models that handle a wide variety of data distributions. It offers the flexibility of selecting specific count data distributions and modeling techniques from a list of seventeen data distributions.

We demonstrate this macro procedure by applying it to a takeover bid data set provided by Jaggia and Thosar (1993) and assess the performance of the count models. Description and the capabilities of the SAS®macro *%countreg* can be found in the appendix A.

Section 2 discusses several count modeling distributions and their properties which are modeled in our *%countreg* macro. In section 3, we describe discuss two models commonly used when data has excess zeros —zero-inflated model and hurdle model. Methods of modeling count data that are flexible in covering various types of data distribution are discussed in section 4. Section 5 provides application results and evaluations of various count models, and concluding remarks are presented in section 6.

2. COUNT MODELING DISTRIBUTIONS AND THEIR PROPERTIES

There are many distributions that deal with count dependent variables. The SAS procedures: GENMOD, GLIMMIX and COUNTREG are limited to the Poisson and negative binomial distributions. Our SAS macro, *%countreg*, is flexible in allowing for a wide variety of data distributions and modeling techniques. Some of these count data distributions are discussed as follows.

2.1 POISSON DISTRIBUTION AND MIX-POISSON DISTRIBUTION

The starting point for cross-section count data analysis is the Poisson regression model. Poisson distribution assumes the equality of the conditional variance and the conditional mean (equidispersion). The density function of a Poisson distribution is:

$$f(y_i \mid \mathbf{x}_i) = \frac{e^{-\mu_i}\mu_i}{y_i!}, \qquad y_i = 0,1,2,....$$

with mean and variance parameters

$$E[y_i \mid \mathbf{x}_i] = V[y_i \mid \mathbf{x}_i] = \mu_i = \exp(\mathbf{x}_i\beta)$$

The Poisson log-likelihood function is

$$L(\beta) = \sum_{i=1}^{n} [y_i \mathbf{x}_i \beta - \mu_i - \log(y_i!)]$$

Equi-dispersion implied by the Poisson distribution is very restrictive. Often the conditional variance of y_i is equal to $\mu_i + \tau\mu_i^2$ and exceeds the conditional mean (overdispersion). Two statistical sources may cause overdispersion: positive contagion and unobserved heterogeneity. To address these issues, mixed-Poisson model includes an unobserved specific effect ε_i into the μ_i parameter; this specific effect can be treated as random or fixed. In the case of random effects, the relationship between $\tilde{\mu}_i$ and μ_i becomes:

$$\tilde{\mu}_i = \exp(\mathbf{x}_i\beta + \varepsilon_i) = \exp(\mathbf{x}_i\beta)\exp(\varepsilon_i) = \mu_i v_i \qquad (2.1.1)$$

The random term ε_i takes into account possible specification errors. The precise form of the distribution of the mixed Poisson model depends upon the specific choice of the probability distribution of v_i. Let $g(v_i)$ be the density of v_i then the conditional density is

$$f(y_i \mid \mathbf{x}_i) = \int \exp(-\tilde{\mu}_i)\tilde{\mu}_i^{y_i} / \Gamma(y_i + 1) g(v_i) dv_i \qquad (2.1.2)$$

2.2 NEGATIVE BINOMIAL (POISSON-GAMMA MIXTURE) DISTRIBUTION

The most common mixed Poisson model is the negative binomial model. This occurs when v_i in (2.1.1 and 2.1.2) is gamma distributed. The most common implementation of the negative binomial is the NB(2) model with probability density function given by:

$$f(y_i \mid \mu_i, \alpha) = \frac{\Gamma(y_i + \alpha^{-1})}{\Gamma(y_i + 1)\Gamma(\alpha^{-1})} \left(\frac{\alpha^{-1}}{\alpha^{-1} + \mu_i}\right)^{\alpha^{-1}} \left(\frac{\mu_i}{\alpha^{-1} + \mu_i}\right)^{y_i} \qquad (2.2.1)$$

When $\alpha = 0$, it is the Poisson distribution. A statistically significant α implies overdispersion. Cameron and Trivedi (1986) considered NB(p) models with mean μ_i and variance function $\mu_i + \alpha\mu_i^p$, p=2 is the standard formulation of the negative binomial model. Models with other values of *p* have the same density as (2.2.1) except that α^{-1} is replaced everywhere by $\alpha^{-1}\mu_i^{2-p}$. The negative binomial log-likelihood function is given by

$$L(\beta, \alpha) = \sum_{i=1}^{n} \left[\log\left(\frac{\Gamma(y_i + \alpha^{-1})}{\Gamma(y_i + 1)\Gamma(\alpha^{-1})}\right) - (y_i + \alpha^{-1})\log(1 + \alpha\mu_i) + y_i \log(\alpha\mu_i) \right]$$

One could also further specify the negative binomial by allowing the α term to vary systemically (generalized negative binomial). This allows the dispersion to vary observation by observation. Let $\alpha = \exp(\mathbf{z}_i\gamma)$, were $\mathbf{z}_i$ could be the all or a subset of the explanatory variables $\mathbf{x}_i$ or completely different variables.

2.3 POISSON- INVERSE GAUSSIAN MIXTURE

The Poisson Inverse Gaussian (PIG) model is an attractive alternative to the negative binomial models when a longer tailed distribution is present. An inverse Gaussian distribution (Folks and Chhikara 1978) for v_i, has density

$$g(v_i) = \left(2\pi\phi v_i^3\right)^{-1/2} e^{-(v-1)^2/2\phi v_i} \quad \text{for } v > 0 \quad \text{where} \quad E[v_i] = 1, V[v_i] = \phi \qquad (2.3.1)$$

The probability generating function for $PIG(\mu, \phi)$ is

$$P(z)=\sum_{0}^{\infty}p(y)z^{y}=\exp\left(1/\phi\mu^{k-1}\left[1-\{1-\phi\mu^{k}(z-1)\}^{1/2}\right]\right) \qquad (2.3.2)$$

Where we have written *p(y)* for *Pr(Y = y)*. Probabilities may be calculated recursively using the established results.

The log-likelihood function is $L(\beta,\phi,k)=\sum_{i=1}^{n}\log p_i(y_i)$ where $p_i(y_i)$ stands for $\Pr(Y=y_i\mid \mathbf{x}_i;\beta,\phi,k)$

Let

$$t_i(y)=(y+1)\frac{p_i(y+1)}{p_i(y)} \quad y=0,1,2,\ldots, \qquad (2.3.3)$$

Manipulation of (2.3.3), and (2.3.4) shows that log-likelihood function can be expressed as:

$$L(\beta,\phi,k)=\sum_{i=1}^{n}\log\left(\frac{1}{y!}\right)+p_i(0)+I(y_i>0)\sum_{j=0}^{y_i-1}\log(t_i(j))$$

2.4 POISSON-NORMAL MIXTURE

This model assumes the heterogeneity term v_i in 2.1.1 as a normally distributed variable with mean zero and standard deviation σ, which we introduce into the model explicitly by standardizing ε_i then, the density is

$$P(y_i\mid \mathbf{x}_i)=\int_{-\infty}^{\infty}\frac{\exp[-\exp(\sigma\varepsilon_i)\mu_i][\exp(\sigma\varepsilon)\mu_i]}{\Gamma(y_i+1)}\phi(\varepsilon_i)d\varepsilon_i$$

Where here and in what follows, $\phi(\varepsilon_i)$ denotes the standard normal density. The log likelihood function is

$$L(\beta,\sigma)=\sum_{i=1}^{n}\ln(P(y_i\mid \mathbf{x}_i))=\sum_{i=1}^{n}\log\left(\int_{-\infty}^{\infty}\frac{\exp[-\exp(\sigma\varepsilon_i)\mu_i][\exp(\sigma\varepsilon)\mu_i]}{\Gamma(y_i+1)}\phi(\varepsilon_i)d\varepsilon\right)$$

2.5 POISSON-SEMIPARAMETRIC (LAGUERRE POISSON)

Poisson-semiparametric model is a flexible model proposed by Gurmu, Rilestone and Stern (1998). It avoids strong parametric assumptions about the distribution of v_i. From 2.1.1 where $\mu_i=\exp(\mathbf{x}_i\beta)$ and $\tilde{\mu}_i=\mu_i v_i$, then from 2.1.2 we have

$$f(y_i\mid \mathbf{x}_i)=\frac{\mu_i^{y_i}}{\Gamma(y_i+1)}\int\exp(-\mu_i v_i)g(v_i)dv_i=\frac{\mu_i^{y_i}}{\Gamma(y_i+1)}M_v^{(y)}(-\mu_i)$$

where $M_v(-\mu_i)$ is the moment generating function for v_i and $M_v^{(y)}(-\mu_i)$ is the y[th]-order derivative of $M_v(-\mu_i)$. The log likelihood function for semiparametric Poisson model is given by

$$L(\beta,\alpha,c_1,c_c,..c_K)=\sum_{i=1}^{n}\left(y_i\log(\mu)_i-\log(\Gamma(y_i+1))+\log\left(M_v^{(y)}(-\mu_i)\right)\right)$$

After a long and tedious derivation it can be shown that

$$M_v^{(y)}(-\mu_i)=\left(1-\frac{-\mu_i}{\lambda}\right)^{-\alpha}(\lambda+\mu_i)^{-y_i}\frac{\Gamma(\alpha)}{\sum_{j=0}^{K}c_j^2}\sum_{j=0}^{K}\sum_{k=0}^{K}\sum_{l=0}^{j}\sum_{m=0}^{k}c_j c_k\left(h_j h_k\right)^{1/2}\binom{j}{l}\binom{k}{m}\frac{\Gamma(\alpha+l+m+y_i)}{\Gamma(\alpha+l)\Gamma(\alpha+m)}\left(-1-\frac{\mu_i}{\lambda}\right)^{-(l+m)}$$

where $h_j=\dfrac{\Gamma(j+\alpha)}{\Gamma(\alpha)\Gamma(j+1)}$ and $\lambda=\dfrac{\Gamma(\alpha)}{\sum_{j=0}^{K}c_j^2}\sum_{j=0}^{K}\sum_{k=0}^{K}\sum_{l=0}^{j}\sum_{m=0}^{k}c_j c_k\left(h_j h_k\right)^{1/2}\binom{j}{l}\binom{k}{m}\dfrac{\Gamma(\alpha+l+m+1)}{\Gamma(\alpha+l)\Gamma(\alpha+m)}(-1)^{-(l+m)}$

This conditional density can generate the Poisson model if $\left(\alpha^{-1}\to 0, c_0=1, c_j=0, \forall j\ge 1\right)$, the geometric model if $\left(\alpha=1, c_0=1, c_j=0, \forall j\ge 1\right)$, and the negative binomial 2 model if $\left(c_0=1, c_j=0, \forall j\ge 1\right)$

2.6 GENERAL EVENT COUNT GEC(K) (KATZ SYSTEM)

Some extensions of the Poisson model that permits both over- and under-dispersion can be obtained by introducing a variance function with an additional parameter. Winkelmann and Zimmermann (1991, 1995), developed a more flexible conditional variance than the NB(*p*). They developed the General Event Count Model (GEC(k)), which is based on a new parameterization of the Katz family. The conditional variance of the GEC(k) model is

$$V(y_i\mid \mathbf{x}_i)=E(y_i\mid \mathbf{x}_i)+(\varsigma-1)E(y_i\mid \mathbf{x}_i)^{k+1}\qquad \varsigma>0$$

Where ς and k, represent the dispersion parameter and the non-linearity in the conditional variance. This more general full parametric specification allows for overdispersion $\varsigma>1$, and underdispersion $0<\varsigma<1$. Furthermore, it encompasses the Poisson model (for $\varsigma=1$), NB(1) (for $\varsigma>1$ and k=0), NB(2) (for $\varsigma>1$ and k=1), and NB(p) for ($\varsigma>1$ and k=p-1) as special cases. By letting $\alpha=(\varsigma-1)$ log likelihood of the GEC(k) is given by

$$L(\beta,\alpha,k)=\sum_{i=1}^{n}\left(I(y_i=0)C_i+I(y_i>0)\sum_{j=1}^{n}\log\left(\mu_i+\alpha(j-1)\mu_i^k/\left(\left(\alpha\mu_i^k+1\right)j\right)\right)\right)$$

Where,

$$C_i=\begin{cases}-\mu_i\log\left(\alpha\mu_i^k+1\right)/\alpha\mu_i^k & \alpha\ge 0\\ -\mu_i\log\left(\alpha\mu_i^k+1\right)/\alpha\mu_i^k-\ln(D_i) & 0<\alpha-1<1;\ \mu_i^k\le 1/\alpha;\ y_i\le \text{int}^*(\xi_i)\\ 0 & \textit{Otherwise}\end{cases}$$

$$D_i=\sum_{m=0}^{\text{int}^*(\xi)}f_{binomial}(m\mid \mu,\alpha,k)$$

2.7 DOUBLE POISSON

The double Poisson distribution (Efron 1986) is obtained as an exponential combination of two Poisson distributions that has a probability density function given by:

$$f(y_i,\mu_i,\phi)=K(\mu_i,\phi)\phi^{1/2}\exp(-\phi\mu_i)\frac{\exp(-y_i)y_i^{y_i}}{y_i!}\left(\frac{e\mu_i}{y_i}\right)^{\phi y_i}$$

Where ϕ is a dispersion parameter, and $K(\mu,\phi)$ is a normalizing constant that ensures the density $f(y,\mu,\phi)$ sums to unity.

$$\frac{1}{K(\mu\phi)}\approx 1+\frac{1-\phi}{12\phi\mu}\left(1+\frac{1}{\phi\mu}\right)$$

This distribution has $E(y_i\mid x_i)=\mu_i$, and $V(y_i\mid x_i)=\mu_i/\phi$, The Poisson model is nested in the double Poisson model for $\phi=1$. The double Poisson model also allows for overdispersion $(\phi<1)$ as well as

underdispersion $(\phi > 1)$. Because the constant $K(\mu,\phi)$ is a source of significant nonlinearity, this term may be suppressed in the maximum likelihood estimation. The log likelihood function for the double Poisson model is.

$$L(\beta,\phi) = \sum_{i=1}^{n} \log(\phi)/2 - \phi\mu_i + y_i(\log(y_i) - 1) - \log(\Gamma(y_i + 1)) + \phi y_i(1 + \log(\mu_i / y_i))$$

2.8 GENERALIZED POISSON

Famoye (1993) proposed the generalized Poisson distribution that can accommodate both over- and under-dispersion. Famoye and Singh (2003) further discussed the inflated generalized Poisson distribution model. This distribution has a probability density function given by

$$f(y_i, \mu_i, \phi) = \left(\frac{\mu_i}{1 + \phi\mu_i}\right)^{y_i} \frac{(1 + \phi y_i)^{y_i - 1}}{y_i!} \exp\left(\frac{-\mu_i(1 + \phi y_i)}{1 + \phi\mu_i}\right)$$

If $\phi = 0$ then the generalized Poisson model reduces to the Poisson model. Also the parameter ϕ is restricted to $1 + \phi\mu_i > 0$ and $1 + \phi y_i > 0$, the model is sometimes called the restricted generalized Poisson model. The mean of generalized Poisson model is $E(y_i \mid x_i) = \mu_i$ and variance $V(y_i \mid x_i) = \mu_i / (1 + \phi\mu_i)^2$ .Clearly, when $\phi > 0$, the variance is overdispersed and when $2/\mu_i < \phi < 0$ the variance is underdispersed. The log likelihood function of the generalized Poisson model is given by

$$L(\beta,\phi) = \sum_{i=1}^{n} y_i \log(\mu_i / 1 + \phi\mu_i) + (y_i - 1)\log(1 + \phi y_i) - \mu_i(1 + \phi y_i)/1 + \phi\mu_i - \log(\Gamma(y_i + 1))$$

2.9 NEYMAN TYPE A (POISSON-STOPPED-SUM-POISSON)

The Neyman type A distribution is a two-parameter distribution which describes discrete data generated by a clustering effect that is commonly observed in the biological science. Such data often has an excessive frequency of zeros and very few counts of one. These distributions can be thought of as compound distributions that involve two processes. The distribution was developed by Neyman (1939) to describe the number of larvae in a field. The distribution is also known as a Poisson-stopped-summed-Poisson distribution. Let λ_i be the average number of clusters of occurrences, and ϕ_i be the average number of occurrences per cluster then the Neyman Type A distribution has probability density function given by

$$f(y_i, \lambda_i, \phi_i) = \frac{\exp(-\lambda_i)\phi_i^{y_i}}{y_i!} \sum_{j=0}^{\infty} \frac{(\lambda_i \exp(-\phi_i))^j}{j!}$$

Where $\lambda_i = \exp(\mathbf{x}_i\beta)$, and $\phi_i = \exp(\mathbf{z}_i\gamma)$ where ***x*** and ***z*** are the explanatory variables. The mean of the response in the Neyman Type A model is $E(y_i \mid x_i) = \lambda_i\phi_i = \mu_i$ and variance $\mu_i + \mu_i^2/\lambda_i$

The log likelihood function of the Neyman Type A model is given by

$$L(\beta,\gamma) = \sum_{i=1}^{n} -\lambda_i + y_i \log(\phi_i) - \log(\Gamma(y_i + 1)) + \log\left(\sum_{j=0}^{\infty} \frac{(\lambda_i \exp(-\phi_i))^j}{j!}\right)$$

2.10 POLYA-AEPPLI (GEOMETRIC POISSON)

The Polya-Aeppli distribution (Johnson, Kotz and Kemp 1992) describes a model where the objects/events occur in clusters, the clusters follow a Poisson distribution with shape parameter λ_i, and the number of objects within a cluster follows a geometric distribution with shape parameter $0 < \tau < 1$. For this reason, this distribution is sometimes referred to as a geometric Poisson distribution with a probability density function given by.

$$f(y_i \mid \lambda_i, \tau) = \begin{cases} \exp(-\lambda_i) & y_i = 0 \\ \exp(-\lambda_i)\tau^{y_i} \sum_{j=1}^{y_i} \binom{y_i - 1}{j-1} \frac{(\lambda_i(1-\tau)/\tau)^j}{j!} & y_i = 1, 2, ... \end{cases}$$

The mean of the response in the Polya-Aeppli model is $E(y_i \mid x_i) = \lambda_i/(1-\tau)$ and variance $\lambda_i(1+\tau)/(1-\tau)^2$ The log likelihood function of the Polya-Aeppli model is given by

$$L(\beta, \tau) = \sum_{i=1}^{n} I(y_i = 0)(-\lambda_i) + I(y_i > 0)\left[-\lambda_i + y\log(\tau) + \log\left(\sum_{j=1}^{y_i} \binom{y_i - 1}{j-1} \frac{(\lambda_i(1-\tau)/\tau)^j}{j!}\right)\right]$$

3. ZERO MODIFIED COUNT MODELS

3.1 HURDLE AND ZERO-INFLATED REGRESSION

Zero modified count models address the situation when the observed data displays a higher fraction of zeros than can be explained through a standard count regression model. There are two ways of handling this situation. First, the hurdle model (Mullahy 1986) or two-part model where the first part is a binary outcome model (logit or probit) and the second part is a truncated count model. The two parts permit the interpretation that positive observations arise form crossing the zero hurdle or threshold. The hurdle model is appealing because it reflects a two-part decision-making process. For example, in the case of the demand for health care, in the first stage, it is up to the patient to decide whether to visit the doctor (contact analysis—probability that the threshold is crossed) and then it is essentially up to the doctor to determine the intensity of the treatment (frequency analysis—truncated count model). The probability density function of the hurdle model is given by

$$f(y_i \mid \mathbf{x}_i) = \begin{cases} F(\mathbf{z}_i\gamma) & y_i = 0 \\ \frac{1 - F(\mathbf{z}_i\gamma)}{1 - f(0 \mid \mathbf{x}_i)} f(y_i \mid \mathbf{x}_i) & y_i > 0 \end{cases}$$

Where $F(\mathbf{z}_i\gamma)$ *Pr(y=0)* is the CDF of the logistic or probit regression selection model with explanatory variables $\mathbf{z}_i$ and parameter estimates γ. The $\mathbf{z}_i$ may be the same explanatory variables as $\mathbf{x}_i$ a subset of them or completely different variables. $f(y_i \mid \mathbf{x}_i)/(1 - f(0 \mid \mathbf{x}_i))$ is the probability density function of a truncated count regression model. The macro *%countreg* allows for any truncated probability density function mentioned in section 2. The log likelihood function for the hurdle model is given by

$$L(\gamma, \beta, \Theta) = \sum_{i=1}^{n} I(y_i = 0)\log(F(\mathbf{z}_i\gamma)) + I(y_i > 0)[\log(1 - F(\mathbf{z}_i\gamma)) + \log(f(y_i \mid \mathbf{x}_i)) - (1 - f(0 \mid \mathbf{x}_i))]$$

Where β is the parameter estimates for the explanatory variables in the truncated count regression model and based on the count data distribution selected, Θ is any additional parameter estimates.

3.1 ZERO-INFLATED MODELS

Another way to model excess zeros in the count modeling is the zero-inflated count models (Lambert 1992). A zero-inflated count model is a special case of a finite mixture model (section 4.2). Zero-inflated model assumes that the zero counts come from two sources, not one source as in the hurdle model. An example of a zero-inflated model may be the number of fishing trips taken over a specified time period. There are people who would never choose to go fishing and others that would but did not go fishing during the sample period. A logit or probit model is used to determine the probability count outcome to be zero. The zero-inflated probability density function is given by

$$f(y_i \mid \mathbf{x}_i) = \begin{cases} F(\mathbf{z}_i\gamma) + (1 - F(\mathbf{z}_i\gamma))f(0 \mid \mathbf{x}_i) & y_i = 0 \\ (1 - F(\mathbf{z}_i\gamma))f(y \mid \mathbf{x}_i) & y_i > 0 \end{cases}$$

Where $F(\mathbf{z}_i\gamma)$ *Pr(y=0)* is the CDF of the logistic or probit regression selection model with explanatory variables $\mathbf{z}_i$ and parameter estimates γ. The $\mathbf{z}_i$ may be the same explanatory variables as $\mathbf{x}_i$ a subset of them or completely different variables. $f(y_i \mid \mathbf{x}_i)$ is the density of a count regression model with explanatory variables $\mathbf{x}_i$. The macro *%countreg* allows for any probability density function mentioned in section 2. The log likelihood function for the zero-inflated model is given by

$$L(\gamma,\beta,\Theta)=\sum_{i=1}^{n} I(y_i=0)\log\left(F(\mathbf{z}_i\gamma)+(1-F(\mathbf{z}_i\gamma))f(0\mid\mathbf{x}_i)\right)+I(y_i>0)\left[\log(1-F(\mathbf{z}_i\gamma))+\log(f(y_i\mid\mathbf{x}_i))\right]$$

Where β is the parameter estimates for the explanatory variables in count regression model and based on the count data distribution selected, Θ is any additional parameter estimates.

For the zero-inflated (tau) model, the $\mathbf{z}_i$ are the same as the $\mathbf{x}_i$ and the parameters in the binary model are assumed to be scalar multiple of the parameters in the count model. Based on these assumptions, the zero-inflated (tau) model reduces the number of parameters to be estimated. The log likelihood function for the zero-inflated (tau) model is

$$L(\beta,\Theta,\tau)=\sum_{i=1}^{n} I(y_i=0)\log\left(F(\tau x_i\gamma)\right)+I(y_i>0)\left[\log(1-F(\tau\mathbf{x}_i\gamma))+\log(f(y_i\mid\mathbf{x}_i))-(1-f(0\mid\mathbf{x}_i))\right]$$

4. FLEXIBLE COUNT MODELING METHODS

4.1 POLYNOMIAL EXPANSION REGRESSION

For under-dispersed data, polynomial expansion regression models (Cameron and Johansson 1997) are used to estimate the parameters. Polynomial expansion count models can handle both under- and over-dispersed data. These models are based on squared polynomial expansions around a baseline density. For underdispersed data the GEC(k) and the generalized Poisson models are limited in restricting the range of the dependent variable. The double Poisson model involves an approximation so that the probabilities do not sum to exactly one. The hurdle model (section 3.1) is another possible model for underdispersed data, however it is not parsimonious- the number of parameters to be estimated is usually doubled. For underdispersed data the series expansion model provides a parsimonious model without restrictions on the range of the dependent variable, while for overdispersed data these models provide an alternative to the negative binomial or other Poisson mixture distributions.

Any count data baseline density could be used for this modeling procedure. However, only the Poisson and the NB(2) densities are supported in the *%countreg* macro. Consider a count variable y_i with baseline density $f(y_i \mid \lambda)$, Define the p[th]-order polynomial as

$$h_p(y_i\mid\mathbf{a})=\sum_{k=0}^{p} a_k y_i^{\,k}$$

Where $\mathbf{a}=(a_0,a_1,a_2,..a_p)$ and $a_0=1$. The density based on a squared polynomial series expansion is

$$g_p(y_i\mid\mathbf{a})=f(y_i\mid\lambda)\frac{h_p^2(y_i\mid\mathbf{a})}{\eta_p(\lambda_i,\mathbf{a})} \qquad (3.1.1)$$

Where $\eta_p(\lambda_i,\mathbf{a})$ is a normalizing constant term that ensures the density $g_p(y_i\mid\lambda_i,\mathbf{a})$ sums to unity, and squaring the polynomial ensures that the density is non-negative.

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

Where $m_r=m_t(\lambda_i)$ denotes the r[th] non-central moment of the baseline density $f(y_i\mid\lambda_i)$.
The log likelihood function for the series expansion polynomial model is given by

$$L(\beta,\mathbf{a})=\sum_{i=1}^{n}\log\left(f\left(y_i\mid\mathbf{x}_i\right)\right)+2\log\left(h_p\left(y_i\mid\mathbf{a}\right)\right)-\eta_p\left(\lambda_i,\mathbf{a}\right)$$

4.2 FINITE AND CONSTRAINED FINITE MIXTURE REGRESSION

The finite mixture models provide a natural and intuitive representation of heterogeneity, usually small number of latent classes, each of which may be regarded as a 'type' or 'group'. In a finite mixture model, a random variable is postulated as a draw from a super-population that is an additive mixture of C distinct populations in proportions $\pi_1,\pi_2,...,\pi_C$. The finite mixture probability density function is given by

$$f\left(y_i\mid\Theta\right)=\sum_{j=1}^{C}\pi_i f_j\left(y_i\mid\Theta_j\right)+\pi_C f_C\left(y_i\mid\Theta\right)\qquad\text{Where}\quad\sum_{j=1}^{C}\pi_j=1,\pi_j>0$$

In general the π_j are unknown and have to be estimated along with all other parameters. $f\left(y_i\mid\Theta\right)$ could be any count distribution. The finite mixture approach is semiparametric: It does not require any distributional assumptions for the mixing variables π_j . The log-likelihood function for the finite mixture model is given by

$$L(\Theta)=\sum_{i=1}^{n}\log\left(f\left(y_i,\mid\Theta\right)\right)$$

Constrained finite mixture model or random intercept model are models in which the j[th] component of the density has intercept parameter θ_j and the slope parameters are restricted to be equal. That is, the subpopulations are assumed to differ randomly only with respect to their location parameter.

5. APPLICATIONS

5.1 DATA

We apply the SAS macro to a takeover bid data set provided by Jaggia and Thosar (1993). The data set include the number of bids received by 126 U.S. firms that were targets of tender offers during the period of 1978-85, and those firms were actually taken over within 52 weeks of the initial offer. The dependent count variable is the number of bids after the initial bid (NUMBIDS) received by the target firm. The independent variables included: defensive actions taken by management of the target firm, firm-specific characteristics, and government intervention. The defensive actions taken by the target firm include: indicator variables for legal defense by lawsuit (LEGALREST), proposed changes in asset structure (REALREST), proposed changed in ownership structure (FINREST), and management invitation for friendly third-party bid (WHITEKNIGHT). The firm-specific characteristics are: bid price divided by price 14 working days before bid (BIDPREM), percentage of stock held by institutions (INSTHOLD), total book value of assets in billions of dollars (SIZE), and book value squared (SIZESQ). The Intervention by federal regulators (REGULATION) is a dummy variable. The summary statistics of all variables are given in Table 5.1 and the frequency distribution of the number of bids is in table 5.2.

Table 5.1 Variable Summary Statistics Distribution

Variable	Mean	Variance	Min	Max
NumBids	1.74	2.05	0.00	10.00
BidPrem	1.35	0.04	0.94	2.07
InstHold	0.25	0.03	0.00	0.90
Size	1.22	9.59	0.02	22.17
Sizesq	11.00	3589.78	0.00	491.46
LegalRest	0.43	0.25	0.00	1.00
RealRest	0.18	0.15	0.00	1.00
FinRest	0.10	0.09	0.00	1.00
Regulation	0.27	0.20	0.00	1.00
WhiteKnight	0.60	0.24	0.00	1.00

Table 5.2 Number of Bids Frequency

Number of Bids	Frequency	Percent of Bids	Cumulative % of Bids
0	9	7.1%	7.1%
1	63	50.0%	57.1%
2	31	24.6%	81.7%
3	12	9.5%	91.3%
4	6	4.8%	96.0%
5	1	0.8%	96.8%
6	2	1.6%	98.4%
7	1	0.8%	99.2%
10	1	0.8%	100.0%

The data has two interesting features—under-dispersion and relatively few zeros (7.1%). There is only a small amount of over-dispersion (20.5/1.74)=1.18, which may disappear as regressors are added. While

the sample average number of bids received after the firs bid is only 1.74 bids, virtually all target firms do receive at least one bid.

5.2 RESULTS

Since the data exhibits under-dispersion, the commonly used negative binomial model is not applicable in this case. For underdispersed data, the Poisson, generalized Poisson, double Poisson, GEC(k) (general event count model), Poisson hurdle, Poisson finite mixture w/ two classes and the Poisson polynomial models are used to estimates the parameters. We apply all the above models to the takeover bid data except for the GEC(k) model due to the convergence problem with the GEC(k) model. Table 5.3 provides the coefficient estimates of all models.

Table 5.3 Takeover Bids: Parameter Estimates Comparisons

Variable	Poisson Regression	Generalized Poisson	Double Poisson	Poisson Hurdle		Poisson Finite Mixture		Poisson Polynomial
				Logit	Poisson	Class 1	Class2	
Intercept	**0.986**	**0.969**	**0.986**	-2.148	1.136	1.092	1.205	0.330
legalrest	**0.260**	**0.265**	**0.260**	-0.971	**0.436**	0.228	0.249	**0.370**
finrest	0.074	0.063	0.074	1.467	0.265	-0.360	0.627	0.084
realrest	-0.196	-0.177	-0.196	**2.723**	-0.004	-0.484	-0.150	-0.230
size	**0.179**	**0.181**	**0.179**	-0.348	**0.238**	0.090	**0.261**	**0.200**
sizesq	**-0.008**	**-0.008**	**-0.008**	-0.013	**-0.010**	-0.004	**-0.011**	**-0.009**
whiteknight	**0.481**	**0.480**	**0.481**	**-1.193**	**0.878**	**0.428**	0.542	**0.833**
bidprem	**-0.678**	**-0.669**	**-0.678**	-0.825	**-1.347**	-0.619	-0.957	**-0.906**
regulation	-0.029	-0.031	-0.029	1.141	-0.057	-0.061	0.053	-0.091
insthold	-0.362	-0.365	-0.362	1.839	-0.661	-0.549	-0.191	-0.605
Alpha		-0.021	**1.422**					
p						0.664		
a1								**3.636**
a2								**-1.341**
a3			1.422**					**0.174**
Log-Likelihood	-185.0	-184.6	-181.5		-159.5		-180.9	-165.7
AIC	389.9	391.2	384.9		359.0		403.7	357.4
BIC	418.3	422.4	416.1		415.7		463.3	394.3

* Bolded coefficients are significant at 10%

$AIC = -2 * LogLikelihood + 2 * (\#\ parameters)$

$BIC = -2Log - Likelihood + Log(\#\ observations)\#\ parameters$

For the double Poisson model, the dispersion parameter $\alpha = 1.42$ is significant at the 5%. Since the variance of the double Poisson distribution $V(y_i \mid \mathrm{x}_i) = \mu_i / \alpha$, with $\alpha > 1$, it implies that number of bids is underdispersed.

The coefficient of WHITEKNIGHT indicates that an invitation for a friendly third-party bid has a positive effect on the number of takeover bids. Another defensive action variable that is statistically significant is the legal defense by lawsuit (LEGALREST), which has a surprising positive sign. It indicates that the legal defense lawsuit invites more bids. The bid-premium has a significant negative effect on the number of takeover bids which is expected. The size of the firm matters, with number of bids increases with the size then decreases with the size squared.

Since the coefficients of the Poisson model and PP3 model are scaled differently, they are not directly comparable. To check the accuracy of the PP3 parameter estimates we compared the mean marginal effects (partial derivative of $E(Y \mid x)$ with respect to x_k) of the PP3 with those from the Poisson model (not reported). There is relatively little difference, with the mean effects for all variables being within 10% of each other.

Using the likelihood ratio test $-2(LL_{Poisson} - LL_{PP3}) = -2(-184.95 - 165.72) = 38.5$ the Poisson model is rejected at 5% when testing against the Poisson Polynomial of order 3 (PP3) model. The Poisson model is also rejected at 5% when tested against the double Poisson and Poisson hurdle model but not rejected against the generalized Poisson or the Poisson finite mixture models.

Two commonly used model selection criteria are the Bayesian information criterion (BIC) and Akaike's information criterion (AIC). Based on the AIC and BIC, the PP3 fits the takeover bids data the best. Although the Poisson hurdle model has better likelihood value it is not parsimonious. The Poisson hurdle model needs to estimate 20 parameters vs. 13 parameters of the PP3 model. Furthermore, one should only use a hurdle model if there is strong theoretical reason for treating zero counts differently from positive counts.

The differences among models lie in their predictive accuracy and aspects of the distribution. Table 5.4 provides the predicted counts of the Poisson, generalized Poisson, double Poisson, Poisson hurdle, Poisson finite mixture with two classes, and the PP3 models along with their Pearson test value. The Pearson's chi-square test is:

$$\chi^2_{Pearson} = \sum_{j=1}^{J} \frac{(Actual - predicted)^2}{predicted}$$ where J is the number of categories

Table 5.4 Takeover Bids: Pearson Chi-Square Goodness of Fit Actual vs. Predicted Counts

Takeover Bids Category	Actual	Poisson Predicted	Generalized Poisson Predicted	Double Poisson Predicted	Poisson Hurdle Predicted	Poisson Finite Mixture Predicted	Poisson Polynomial Predicted
0	9	27	26	18	9	28	9
1	63	38	38	42	62	38	61
2	31	29	30	33	30	29	34
3	12	17	18	18	14	16	12
4	6	9	9	8	6	8	4
5+	5	7	5	7	5	7	6
Chi-Square		32.0	30.4	18.7	0.3	31.9	1.1

Clearly, Table 5.4 indicates that the Poisson model significantly over predicts the number of zeros and under predicts the number of ones. The Pearson chi-square for the Poisson model is 32.01 compared to a $\chi^2(5)$ critical value of 9.24 at 5%. The Poisson model is rejected.

Comparing to other models, both the Poisson hurdle and PP3 models provide superior predicted results. Both models are accepted based on Pearson chi-square test. These results indicate that Poisson hurdle and PP3 models are better than Poisson model when the count data is under-dispersed. Alternative count models should be considered when the underlining data distribution is different from the Poisson distribution.

6. CONCLUSION

This paper provides a flexible SAS macro program that specifies a variety of count data regression models. Although the SAS software provides some procedures for count regression models (e.g. GENMOD, GLIMMIX, COUNTREG), they are restricted to the Poisson and negative binomial distributions. The proposed macro, *%countreg,* is capable of handling many types of data distributions in addition to the Poisson and negative binomial distributions. Furthermore, the SAS macro program is flexible in dealing with count data that exhibits: excessive zeros, under-dispersion, truncation, or censoring.

The various types of data distribution and their respective log likelihood functions were discussed in section 2. We further discussed count models that deal with excessive zero counts and models that are flexible enough to handle both over- and under-dispersed data. Our empirical results confirm that not all data sets can be best analyzed with the Poisson or negative binomial distributions. Depending on the data distribution, different count regression models may provide the best results. For the takeover bids data, that exhibited few zeros and was under-dispersed the Poisson model perform poorly. The Poisson hurdle and Poisson Polynomial of order 3 models were the ones with the best predicted counts. Our

macro provides a flexible program that takes into account a variety of distributions. Its flexibility allows one to use it to analyze any type of count data.

APPENDIX A

```
%macro countreg(indata=,method=,nb=,depend=,indep=,zindep=,gindep=,censor=,
                lcensor=,rcensor=,ctype=,trunc=,ltrunc=,rtrunc=,ttype=,hurdle=,
                zip=,zipt=,poly=,order=,nclass=,summary=,gofcut=,vuong=,vdata1=,
                vdata2);
*-------------------------------------------------------------------------------*
|  Macro countreg performs count regression modeling using the Proc NLMIXED     |
|  procedure. The macro incorporates many different count distributions that    |
|  you can choose from. The inputs into the macro include:                      |
*-------------------------------------------------------------------------------*
|                                                                               |
|  indata =  The name of the SAS data set that contains the count data          |
|  method =  Method used to estimate the count data regression model.           |
|            Choices of methods include:                                        |
|            1  Poisson                                                        |
|            2  Generalized Poisson                                            |
|            3  Poisson-Normal Mixture                                         |
|            4  Poisson-Inverse Gaussian Mixture(k)                            |
|            5  Double Poisson                                                 |
|            6  GEC(k) Generalized Event Count                                 |
|            7  Negative Binomial(1 or 2) Poisson-Gamma Mixture                |
|            8  Generalized Negative Binomial(p)                               |
|            9  Poisson Finite Mixture                                         |
|            10 Negative Binomial(1 or 2) Finite Mixture                       |
|            11 Poisson Constrained Finite Mixture                            |
|            12 Negative Binomial(1 or 2) Constrained Finite Mixture           |
|            13 Poisson Semi-Parametric(k)                                     |
|            14 Poisson Polynomial Expansion(k)                                |
|            15 Negative Binomial(1 or 2) Polynomial Expansion                 |
|            16 Polya-Aeppli                                                   |
|            17 Neyman Type A                                                  |
|  nb =      The order of the Negative Binomial(1 or 2)                         |
|  depend =  Name of the dependent variable                                     |
|  indep  =  Name of the explanatory variables                                  |
|  zindep =  Name of the independent variables that are used in the hurdle      |
|            or the zero-inflated model (selection equation)                    |
|  gindep =  Name of the independent variables that are used in the alpha       |
|            parameter in the generalized negative binomial or the variables    |
|            used for the phi parameter in the Neyman Type A model              |
|  censor =  If the data is censored(0=No,1=Yes)                                |
|  lcensor=  If data is left censored what value is it censored at              |
|  rcensor=  If data is right censored what value is it censored at             |
|  ctype  =  What direction is the data censored (Left, Right or Both)          |
|  trunc  =  If the data is truncated(0=No,1=Yes                                |
|  ltrunc =  If data is left truncated what value is it truncated at            |
|  rtrunc =  If data is right truncated what value is it truncated at           |
|  ttype  =  What direction is the data truncated (Left, Right or Both)         |
|  hurdle =  If you want to estimate a hurdle model(0=No,1=Yes)                 |
|  zip    =  If you want to estimate a zero-inflated model(0=No,1=Yes)          |
|  zipt   =  If you want to estimate a zero-inflated(tau) model(0=No,1=Yes)     |
|  poly   =  Order of the Poisson-Semi-Parametric model                         |
|  order  =  Order of the (Poisson or NB) series expansion model                |
|  nclass =  The number of latent classes that are used for the Poisson and     |
|            Negative Binomial Finite Mixture and Constrained Finite Mixture   |
|  summary=  Summary statistics on the dependent and independent variables      |
|            (0=No,1=Yes)                                                       |
|  gofcut =  Number of categories-1 you want for the Pearson Chi-Square test    |
|  vuong  =  Test two non-nested models with the Voung test (0=No,1=Yes)        |
|  vdata1 =  data set name that contains the log-likelihood values for the      |
```

```
|            first model you want to use with the Vuong test.                |
|  vdata2 =  data set name that contains the log-likelihood values for the   |
|            second model you want to use with the Vuong test.               |
*-------------------------------------------------------------------------
------*;
```

REFERENCES

Cameron, A. C. and Johansson Per (1997) Count Data Regression Using Series Expansions: With Applications, *Journal of Applied Econometrics*, Vol. 12, No.3, 203-223.

Cameron, A. C. and Trivedi, P. K. (1996), Count Data Models for Financial Data, Handbook of Statistics, Vol. 14,
Statistical Methods in Finance, 363-392, Amsterdam, North-Holland.

Cameron, A. C. and Trivedi, Pravin K. (1998), Regression Analysis of Count Data, *Econometric Society Monographs, No. 30*, Cambridge University Press.

Czado, Claudia, Vinzenz Erhardt, Aleksey Min, Stefan Wagner. (2007). Zero-inflated Generalized Poisson Models with Regression Effects on the Mean, Dispersion and Zero-inflation Level Applied to Patent Outsourcing rates. *Statistical Modelling*, 7(2), 125-153.

Dean, C., Lawless, J.F., Willmot, (1989), A Mixed Poisson-Inverse Gaussian Regression Model, *The Canadian Journal of Statistics,* 17(2), 171-181

Deb, Partha and Trivedi, Pravin K. (1997), Demand for Medical Care By The Elderly: A Finite Mixture Approach, *Journal of Applied Econometrics*, 12, 313-336.

Efron, B. (1986), Double Exponential Families and Their Use in Generalized Linear Regressions, *Journal of the American Statistical Association, 81, 709-721.*

Famoye, F (1993) Restricted Generalized Poisson Regression Model, *Communications in Statistics, Theory and Methods*, **22**(5), 1335–54.

Famoye, F and Singh KP (2003) On Inflated Generalized Poisson Regression Models, *Advances and Applications in Statistics*, **3**(2), 145–58.

Folks, J.L., and Chhikara, R.S. (1978). The inverse Gaussian distribution and its statistical application-a review. *J. Roy. Statist. Soc. Ser. B,* 40, 263-289.

Greene, William (2007), Functional Form and Heterogeneity in Models for Count Data, Working Paper, *Department of Economics, Stern School of Business, New York University.*

Guikema, Seth D. and Jeremy P. Goffelt (2008), Flexible Count Data Regression Model for Risk Analysis, Risk Analysis, 28(1), 213-223.

Gurmu, Shiferaw, Rilestone, Paul and Stern, Steven, (1999), Semiparametric Estimation of Count Regression Models, *Journal of Econometrics 88, 123-150.*

Hellström, Jörgen and Jonas Nordström (2008), A Count Data Model with Endogenous Household Specific Censoring: the Number of Nights to Stay, *Empirical Economics*, Vol. 35, No. 1, 179-193.

Jaggia, S. and Thosar, S., (1993), Multiple Bids as a Consequence of Target Management Resistance: A Count Data Approach, *Review of Quantitative Finance and Accounting*, December, 447-57.

Johnson, Norman L., Kotz, Samuel, Kemp, Adrienne W., Univariate Distributions, *2nd edition, John Wiley & Sons, New York 1992.*

King, Gary (1989), Variance Specification in Event Count Models: From Restrictive Assumptions to a Generalized Estimator. *American Journal of Political Science*, Vol. 33 No.3, 762-84.

Lambert, D (1992), Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing, *Technometrics, Vol. 34, No. 1, 1 – 14.*

Liu, WenSui and Jimmy Cela (2008), Count Data Models in SAS, Proceedings SAS Global Forum 2008, paper 371-2008.

Long, J. Scott, (1997), Regression Models for Categorical and Limited Dependent Variables, *Advanced Quantitative Techniques in the Social Sciences Series 7*, SAGE Publications

Min, Yongyi , Agresti, Alan (2002), Modeling Nonnegative Data with Clumping at Zero: A Survey, JIRSS Vol.1, Nos. 1-2, 7-33.

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

Nelson, Jon P. and Douglas Young (2008), Effects of youth, price, and audience size on alcohol advertising in magazines, Health Economics, 17(4), 551-556.
Neyman, J. (1939), A New Class of Contagious Distributions Applicable in Entomology and Bacteriology, *Annals of Mathematical Statistics, 10, 35-57.*
Tin, Adrienne (2008), Modeling Zero-Inflated Count Data with Underdispersion and Overdispersion, Proceedings SAS Global Forum, Paper 372-2008.
Winkelmann, R. and K. F. Zimmermann (1991), A New Approach for Modeling Economic Count Data, *Economics Letters*, Vol. 37,139-143.
Winkelmann, R. and K. F. Zimmermann (1995), Recent Development in Count Data Modeling: Theory and Application, *Journal of Economic Surveys*, Vol. 9, 1-24.
Yang, Chih-Hai (2007), What factors inspire the high entry flow in Taiwan's manufacturing industries-A count entry model approach, Applied Economics. Vol. 39, 1817-1831.

CONTACT INFORMATION

Nan-Ting Chou
University of Louisville,
Economics Department, College of Business
2301 South Third Street
Louisville, KY 40208
ntchou01@louisville.edu

David Steenhard, Senior Statistician
LexisNexis
9443 Springboro Pike
Dayton, OH 45342
Email: david.steenhard@lexisnexis.com