Journal of Data Science **4**(2006), 117-130

Zero-Inflated Generalized Poisson Regression Model with an Application to Domestic Violence Data

Felix Famoye[1] and Karan P. Singh[2]
[1] *Central Michigan University and* [2] *UNT Health Science Center*

Abstract: The generalized Poisson regression model has been used to model dispersed count data. It is a good competitor to the negative binomial regression model when the count data is over-dispersed. Zero-inflated Poisson and zero-inflated negative binomial regression models have been proposed for the situations where the data generating process results into too many zeros. In this paper, we propose a zero-inflated generalized Poisson (ZIGP) regression model to model domestic violence data with too many zeros. Estimation of the model parameters using the method of maximum likelihood is provided. A score test is presented to test whether the number of zeros is too large for the generalized Poisson model to adequately fit the domestic violence data.

Key words: Count data, parameter estimation, score test, zero-inflation.

1. Introduction

The generalized Poisson regression (GPR) model proposed by Consul and Famoye (1992) and Famoye (1993) is used to model count data that are affected by a number of known predictor variables. The model is based upon the generalized Poisson distribution which had been extensively studied by researchers. The reader is referred to Consul (1989) and the references therein for more details. The GPR model has been used to model a household fertility data set (Wang and Famoye, 1997) and to model injury data (Wulu *et al.*, 2002).

Count data with too many zeros are common in a number of applications. Ridout *et al.* (1998) cited examples of data with too many zeros from various disciplines including agriculture, econometrics, patent applications, species abundance, medicine, and use of recreational facilities. Several models have been proposed to handle count data with too many zeros than expected: Lambert (1992) described the zero-inflated Poisson (ZIP) regression models with an application to defects in manufacturing; Hall (2000) described the zero-inflated binomial (ZIB) regression model and incorporated random effects into ZIP and ZIB models; and Lee *et al.* (2001) generalized the ZIP model to accommodate the extent

of individual exposure. Other models in the literature include the hurdle model (Mullahy, 1986), the two-part model (Heibron, 1994), and the semi-parametric model (Gurmu, 1997). Details of these models can be found in Ridout *et al.* (1998) and additional references on ZIP models can be found in Bohning *et al.* (1999).

A feature of many count datasets is the joint presence of excess zero observations and the long right tails, both relative to the Poisson assumption, Gurmu and Trivedi (1996). Both features may be accounted for by over-dispersion in the data. The excess zeros can occur as a result of clustering. Over-dispersion has the tendency to increase the proportion of zeros and whenever there are too many zeros relative to Poisson assumption, the negative binomial regression and the generalized Poisson regression tend to improve the fit of the data. For a better fit, an over-dispersed model that incorporates excess zeros should serve as an alternate. This point was illustrated by Gurmu and Trivedi (1996) who found that the negative binomial hurdles model, which allows for over-dispersion and also accommodates the presence of excess zeros, is more appropriate among all the models they considered. Also, Ridout *et al.* (1998) considered various ZIP regression models for an Apple shoot propagation data. They concluded that the ZIP models were inadequate for the data as there was still evidence of over dispersion. They went on to fit zero-inflated negative binomial models to the data.

Gupta *et al.* (1996) studied the zero-adjusted generalized Poisson distribution. They estimated the model parameters by the method of maximum likelihood. They studied the effect of not using adjusted (inflated or deflated) model when the occurrence of zero differs from what is expected. They showed that more errors are committed for small values of the count if adjustment is ignored. They noted that the zero-adjusted generalized Poisson distribution fitted very well the fetal movement data and the death notice data of London times. In this paper, we extend their work to a more general situation where the count dependent variable is affected by some covariates.

In our research work, we have seen cases where the ZIP models were inadequate and the zero-inflated negative binomial regression model could not be fitted to the data sets. The major problem in these cases was that the iterative technique to estimate the parameters of zero-inflated negative binomial regression model failed to converge. This observation is similar to the one made by Lambert (1992) and we quote her remark here: "Of course, inflating a negative binomial model with 'perfect zeros' might provide an even better model for the printed-wiring-board data than ZIP regression does. Such a model was not successfully fit to these data, however." This realization motivated us to develop a zero-inflated generalized Poisson regression model for modeling over-dispersed

count data with too many zeros.

The rest of the paper is organized as follows: In section 2, we describe the domestic violence data. We develop the zero-inflated generalized Poisson (ZIGP) regression model in section 3. Estimation of its parameters via the maximum likelihood method is presented in section 4. A score statistic for testing zero inflation in generalized Poisson model is proposed in section 5. The results of applying ZIGP regression to model the number of domestic violence are presented in section 6. In section 6, we also provide some concluding remarks.

2. Description of Domestic Violence Data

In 1989, the Portland Police Bureau in collaboration with the Family Violence Intervention Steering Committee of Multnomah County in Oregon developed a plan to reduce domestic violence in Portland. A special police unit called Domestic Violence Reduction Unit was created for accomplishing two goals: (i) Increasing the sanctions for batterers, and (ii) Empowering victims. A study was designed and data were collected from official records on batterers and from surveys on victims for 1996-1997. For more details the reader is referred to Annette *et al*[1]. (1998), ICPSR 3353. We consider Survey Part 12 data set for illustrating the usefulness of the ZIGP regression model. Data in Part 12 (Wave 2 Victim Interview Data) represent victims' responses to the second wave of interviews, conducted approximately six months after the study case victimization occurred. The descriptive statistics for the variables are given in Table 1.

The variable, violence, is the number of violent behavior of batterer towards victim. In general, an incident of violence may be classified as 'minor' or 'severe'. In this paper we develop a violence index by summing the responses to questions 53 through 62. These questions deal with severe form of violence, for example, threw something; pushed, grabbed, or shoved; slapped; kicked, bit, or hit with a fist; hit or tried to hit with something; and beat up. The independent variables used in the regression models are level of education, employment status, level of income, having family interaction, belonging to a club, and having drug problem. Each of these variables was measured for both victim and batterer. The level of education (from 1 to 3) and income level (from 1 to 5) are ordinal. Other variables are dichotomous with 1 (yes) and 0 (no). After excluding the cases having missing information we have 214 cases.

[1]Annette, J., Fountain, R., Feyethern, W., and Friedman, S. (1998). Portland (Oregon). Domestic Violence Experiment, 1996-97 [computer file], ICPSR Version, Portland, OR: Portland State University (Producer). Ann Arbor, MI: Inter-university Consortium for Political and Social Research (distributor), 2002.

Table 1: Descriptive statistics for the variables

Variable	Description	Mean ± SD	Proportion of 1's
Edu_v	Education level, victim	2.2897± 0.7507	
Edu_b	Education level, batterer	2.0654± 0.7785	
Emp_v	Full time employment, victim		0.5047
Emp_b	Full time employment, batterer		0.6589
Inc_v	Income level, victim	2.5654±1.3083	
Inc_b	Income level, batterer	3.0701±1.4727	
Fam_v	Interact with family, victim		0.8224
Fam_b	Interact with family, batterer		0.7196
Club_v	Belong to a club, victim		0.2710
Club_b	Belong to a club, batterer		0.1916
Drug_v	Have drug problem, victim		0.1355
Drug_b	Have drug problem, batterer		0.6215
Violence	Number of domestic violence	4.2056±10.6014	

SD = standard deviation

3. Zero-Inflated Generalized Poisson Regression Model

Let the response variable $y_i, i = 1, 2, \ldots, n$, be the number of violent behavior of batterer towards victim. The generalized Poisson regression (GPR) model $f(\mu_i, \alpha; y_i)$, is given by

$$f(\mu_i, \alpha, y_i) = \left(\frac{\mu_i}{1+\alpha\mu_i}\right)^{y_i} \frac{(1+\alpha y_i)^{y_i-1}}{y_i!} \exp\left[\frac{-\mu_i(1+\alpha y_i)}{1+\alpha\mu_i}\right], \tag{3.1}$$

$y_i = 0, 1, 2, \ldots$; where $\mu_i = \mu_i(x_i) = \exp\left(\sum x_{ij}\beta_j\right), x_i = (x_{i1} = 1, x_{i2}, \ldots, x_{ik})$ is the i-th row of covariate matrix $\mathbf{X}$, and $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_k)$ are unknown k-dimensional column vector of parameters. The mean of y_i is given by $\mu_i(x_i)$ and the variance of y_i is given by $V(y_i \,|\, x_i) = \mu_i(1+\alpha\mu_i)^2$. In a more general setting, the mean of y_i can be written as $E(y_i \,|\, x_i) = \mu_i(x_i) = c_i\Lambda(x_i, \boldsymbol{\beta})$ where $\Lambda(x_i, \boldsymbol{\beta})$ is a known function of x_i and $\boldsymbol{\beta}$, and c_i is a measure of exposure. The link function $\Lambda(x_i, \boldsymbol{\beta})$ is differentiable with respect to $\boldsymbol{\beta}$. The GPR model in (3.1) is a natural extension of the Poisson regression model given by Frome *et al.* (1973). In model (3.1), α is called the dispersion parameter. When $\alpha = 0$, the probability model in (3.1) reduces to the Poisson regression model and this is a case of equi-dispersion. When $\alpha > 0$, the GPR model in (3.1) represents count data with over-dispersion. When $\alpha < 0$, the GPR model represents count data with under-dispersion.

A zero-inflated generalized Poisson (ZIGP) regression model is defined as

$$\begin{aligned} P(Y = y_i \,|\, x_i, z_i) &= \varphi_i + (1-\varphi_i) f(\mu_i, \alpha; 0), \quad y_i = 0 \\ &= (1-\varphi_i) f(\mu_i, \alpha; 0), \qquad\quad y_i > 0 \end{aligned} \tag{3.2}$$

where $f(\mu_i, \alpha; y_i), y_i = 0, 1, 2, \ldots$ is the GPR model in (3.1) and $0 < \varphi_i < 1$. In (3.2), the functions $\mu_i = \mu_i(x_i)$ and $\varphi_i = \varphi_i(z_i)$ satisfy $\log(\mu_i) = \sum_{j=1}^{k} x_{ij}\beta_j$ and $\text{logit}(\varphi_i) = \log(\varphi_i[1-\varphi_i])^{-1} = \sum_{j=1}^{m} z_{ij}\delta_j$ where $z_i = (z_{i1} = 1, z_{i2}, \ldots, z_{im})$ is the i-th row of covariate matrix $\mathbf{Z}$ and $\boldsymbol{\delta} = (\delta_1, \delta_2, \ldots, \delta_m)$ are unknown m-dimensional column vector of parameters. In this set up, the non-negative functions φ_i and μ_i are, respectively, modeled via logit and log link functions. Both are linear functions of some covariates. Other appropriate link functions that can allow φ_i being negative, in the terminology of generalized linear models, may be used.

The mean and variance of the ZIGP model in (3.2) are given, respectively, by

$$E(y_i \,|\, x_i) = (1-\varphi_i)\mu_i(x_i) \tag{3.3}$$

and

$$\begin{aligned} V(y_i \,|\, x_i) &= (1-\varphi_i)[\mu_i^2 + \mu_i(1+\alpha\mu_i)^2] - (1-\varphi_i)^2\mu_i^2 \\ &= E(y \,|\, x_i)[(1+\alpha\mu_i)^2 + \varphi_i\mu_i]. \end{aligned} \tag{3.4}$$

From (3.4), the distribution of y_i exhibits over-dispersion when $\varphi_i > 0$. The model in (3.2) reduces to the GPR model when $\varphi_i = 0$. It reduces to the ZIP model given by Lambert (1992) when $\alpha = 0$. For positive values of φ_i, it represents the zero-inflated generalized Poisson regression model. When φ_i is allowed to be negative, it represents zero-deflated generalized Poisson regression model. However, zero-deflation rarely occurs in practice.

The covariates affecting φ_i and μ_i may or may not be the same. If the same covariates affect φ_i and μ_i, we can write φ_i as a function of μ_i to obtain

$$\log(\mu_i) = \sum_{j=1}^{k} x_{ij}\beta_j \quad \text{and} \quad \text{logit}(\varphi_i) = \log\left(\frac{\varphi_i}{1-\varphi_i}\right) = -\tau \sum_{j=1}^{k} x_{ij}\beta_j \tag{3.5}$$

The ZIGP regression model with logit link for φ_i and log link for μ_i as defined in (3.5) will be denoted by ZIGP(τ). When $\tau > 0$, the zero state becomes less likely and when $\tau < 0$, excess zeros become more likely. When $\alpha = 0$, the ZIGP(τ) reduces to the ZIP(τ) defined by Lambert (1992).

If y_i are independent random variables having a zero-inflated generalized Poisson distribution, the zeros are assumed to occur in two distinct states. The only occurrences in the first state are zeros which occur with probability φ_i. These are

referred to as 'structural' zeros. The second state occurs with probability $(1-\varphi_i)$ and leads to a generalized Poisson distribution with parameters α and μ_i. The zeros from the second state, i.e. from the generalized Poisson distribution, are called 'sampling' zeros. The two state process leads to a two-component mixture distribution with probability mass function given in (3.2).

In many applications, there is little prior information about how φ_i is related to μ_i (Lambert, 1992). Depending on the data generating process, one can think of a situation in which both φ_i and μ_i depend on some covariates and a situation in which this is not the case. Consider a data set on adults, aged 65-70 years. We may count how many accidents adult aged 65-70 years had while driving during the past five years. A large number of these adults may not have any accidents as they did not drive in the past five years (as opposed to being careful drivers with no accidents). We may be able to model whether an adult drove (during the past five years) depending on a number of covariates related to whether the adults drove or not. We may also model how many accidents an adult had depending on a number of covariates having to do with his/her driving. Thus, we can think of different covariates that will affect φ_i and μ_i. On the other hand, suppose the data is collected from adult drivers who drove through out the past five years. The data could still have too many zeros. In this situation, the ZIGP(τ) model will be more appropriate. Alternatively, one can use the ZIGP regression model with $\mu_i = \mu_i(x_i)$ and $\varphi_i = \varphi_i(z_{i1})$.

4. Parameter Estimation

When φ_i and μ_i are related, the log-likelihood for ZIGP(τ) regression model is given by

$$\begin{aligned}\log(L_\tau) &= -\sum_{i=1}^{n}\log(1+\mu_i^{-\tau}) + \sum_{y_i=0}\log(\mu_i^{-\tau}+\exp[-\mu_i/(1+\alpha\mu_i)]) \\ &\quad + \sum_{y_i>0}\{y_i\log[\mu_i/(1+\alpha\mu_i)] + (y_i-1)\log(1+\alpha y_i) - \log(y_i\,!) \\ &\quad - \mu_i(1+\alpha y_i)/(1+\alpha\mu_i)\}. \qquad (4.1)\end{aligned}$$

In the rest of the paper, we shall use $\xi_i = \mu_i[1+\alpha\mu_i]^{-1}$, $\eta_i = \mu_i^\tau\exp(-\xi_i)$, and $\upsilon_i = \exp(-\xi_i)$. On differentiating (4.1), the likelihood equations are given by

$$\frac{\partial\log(L_\tau)}{\partial\tau} = \sum_{i=1}^{n}\frac{\log(\mu_i)}{1+\mu_i^\tau} - \sum_{y_i=0}\frac{\log(\mu_i)}{1+\eta_i}, \qquad (4.2)$$

$$\frac{\partial\log(L_\tau)}{\partial\beta_r} = \sum_{i=1}^{n}\frac{\tau x_{ir}}{1+\mu_i^\tau} - \sum_{y_i=0}\frac{(\tau+\xi_i^2\eta_i/\mu_i)x_{ir}}{1+\eta_i} + \sum_{y_i>0}\frac{(y_i-\mu_i)x_{ir}}{(1+\alpha\mu_i)^2}, \qquad (4.3)$$

$r = 1, 2, \ldots, k$; and

$$\frac{\partial \log(L_\tau)}{\partial \alpha} = \sum_{y_i=0} \frac{\xi_i^2 \eta_i}{1+\eta_i} + \sum_{y_i>0} \left\{ -\xi_i y_i + \frac{y_i(y_i-1)}{1+\alpha y_i} - \frac{\mu_i(y_i-\mu_i)}{(1+\alpha\mu_i)^2} \right\}. \tag{4.4}$$

The parameters $\tau, \boldsymbol{\beta}$, and α are estimated by the Newton-Raphson algorithm. To fit this model, we first fit the GPR model in (3.1) and the final estimates from GPR are used as the initial values for ZIGP(τ). The final estimate of τ in ZIP(τ) can be taken as an initial guess for τ.

When φ_i and μ_i are not related, the log-likelihood for ZIGP regression model is given by

$$\begin{aligned} \log(L) &= -\sum_{i=1}^{n} \log(1+\omega_i) + \sum_{y_i=0} \log(\omega_i + \upsilon_i) \\ &\quad + \sum_{y_i>0} \{ y_i \log(\xi_i) + (y_i - 1)\log(1+\alpha y_i) \\ &\quad - \log(y_i\,!) - \xi_i(1+\alpha y_i) \}, \end{aligned} \tag{4.5}$$

where $\omega_i = \varphi_i/(1-\varphi_i) = \exp(\sum_{j=1}^{m} z_{ij}\delta_j)$. By using a similar argument as in Lambert (1992), the maximum likelihood via the EM algorithm can be exploited to estimate the parameters $\boldsymbol{\beta}, \boldsymbol{\delta}$ and α in the above log-likelihood function. By differentiating (4.5), we obtain the likelihood equations as follows:

$$\frac{\partial \log(L)}{\partial \beta_r} = -\sum_{y_i=0} \frac{\upsilon_i \xi_i^2 x_{ir}}{(\omega_i + \upsilon_i)\mu_i} + \sum_{y_i>0} \frac{(y_i - \mu_i)x_{ir}}{(1+\alpha\mu_i)^2}, \quad r = 1, 2, \ldots, k \tag{4.6}$$

$$\frac{\partial \log(L)}{\partial \delta_t} = -\sum_{y_i=0} \frac{\omega_i z_{it}}{1+\omega_i} + \sum_{y_i=0} \frac{\omega_i z_{it}}{\omega_i + \upsilon_i}, \quad t = 1, 2, \ldots, \ldots, m, \tag{4.7}$$

and

$$\frac{\partial \log(L)}{\partial \alpha} = \sum_{y_i=0} \frac{\upsilon_i \xi_i^2}{\omega_i + \upsilon_i} + \sum_{y_i>0} \left\{ -\xi_i y_i + \frac{y_i(y_i-1)}{1+\alpha y_i} - \frac{\mu_i(y_i-\mu_i)}{(1+\alpha\mu_i)^2} \right\}. \tag{4.8}$$

Based on the asymptotic normality of the maximum likelihood estimator $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\delta}}, \hat{\alpha})$, inferences on the regression coefficients and the dispersion parameter can be made.

The Newton-Raphson algorithm may be used to find the solutions of the likelihood equations for both ZIGP(τ) and ZIGP. In the application in section 6, we have used the SPLUS function '*nlminb*', to obtain the maximum likelihood estimates. In all of the examples we have considered, the algorithm converged in less than 20 iterations.

5. Score Test for Zero Inflation in Generalized Poisson Model

A score test is proposed to test whether the number of zeros is too large for a generalized Poisson model to adequately fit the data. The reader is referred to Cox and Hinkley (1974) for a discussion of the score test. The score statistics will be obtained for a case with no covariates and for a case with covariates.

5.1 The Case with no Covariates

The inflated generalized Poisson distribution can be obtained from the model in (3.2) if the mean $\mu_i(=\mu)$ and the probability $\varphi_i(=\varphi)$ are constants. Consider the case where there are n observations, among them n_0 zero, and no covariates. By using $\theta = \varphi(1-\varphi)^{-1}$, the log-likelihood function for the zero-inflated generalized Poisson distribution can be written as

$$\begin{aligned}\log(L_*) &= -\sum_{i=1}^{n}\log(1+\theta) + \sum_{y_i=0}\log(\theta+\exp(-\mu/(1+\alpha\mu)) \\ &\quad + \sum_{y_i>0}\{y_i\log[\mu/(1+\alpha\mu)] + (y_i-1)\log(1+\alpha y_i) \\ &\quad - \log(y_i\,!) - \mu(1+\alpha y_i)/(1+\alpha\mu)\} \qquad (5.1)\end{aligned}$$

The score function $U(\mu,\alpha,0)$ and the expected information matrix $I(\mu,\alpha,0)$ can be calculated from (5.1). The score statistic for testing $\theta = 0$ is

$$S(\hat{\mu},\hat{\alpha}) = S(\hat{\mu},\hat{\alpha},0) = U'(\hat{\mu},\hat{\alpha},0)[I(\hat{\mu},\hat{\alpha},0]^{-1}U(\hat{\mu},\hat{\alpha},0). \qquad (5.2)$$

The elements of the score function $U(\alpha,\alpha,0)$ are

$$\frac{\partial\log(L_*)}{\partial\mu} = 0 \qquad (5.3)$$

$$\frac{\partial\log(L_*)}{\partial\alpha} = \sum_{i=1}^{n}\left\{\frac{y_i(y_i-1)}{1+\alpha y_i} - \frac{\mu y_i}{1+\alpha\mu} - \frac{\mu(y_i-\mu)}{(1+\alpha\mu)^2}\right\}, \qquad (5.4)$$

and

$$\frac{\partial\log(L_*)}{\partial\theta} = n_0 f_0 - n, \qquad (5.5)$$

where $f_0 = \exp[\mu/(1+\alpha\mu)]$. The entries in the 3×3 symmetric matrix $I(\mu,\alpha,0)$ are given by $I_{11} = n\mu^{-1}(1+\alpha\mu)^{-2}$; $I_{12} = I_{21} = 0$; $I_{13} = I_{31} = -n(1+\alpha\mu)^{-2}$; $I_{22} = 2n\mu^2(1+2\alpha)^{-1}(1+\alpha\mu)^{-2}$; $I_{23} = I_{32} = n\mu^2(1+\alpha\mu)^{-2}$; and $I_{33} = n[\exp(\mu/(1+$

$\alpha\mu)) - 1]$. On using these values in (5.2), we obtain

$$S(\hat{\mu}, \hat{\alpha}, 0) = \left[\frac{(1+2\hat{\alpha})(1+\hat{\alpha}\hat{\mu})^2}{2n\hat{\mu}^2} + \frac{(1+2\hat{\alpha})^2}{4b}\right] c^2 - \frac{(1+2\hat{\alpha})(n_0 f_0 - n)c}{b} + \frac{(n_0 f_0 - n)^2}{b}, \quad (5.6)$$

where

$$b = n(f_0 - 1) - \frac{n\hat{\mu}}{(1+\hat{\alpha}\hat{\mu})^2} - \frac{n(.5+\hat{\alpha})\hat{\mu}^2}{(1+\hat{\alpha}\hat{\mu})^2} \quad \text{and}$$

$$c = \sum_{i=1}^{n} \left\{ \frac{y_i(y_i-1)}{1+\hat{\alpha}y_i} - \frac{\hat{\mu}y_i}{1+\hat{\alpha}\hat{\mu}} - \frac{\hat{\mu}(y_i-\hat{\mu})}{(1+\hat{\alpha}\hat{\mu})^2} \right\}.$$

Table 2: Percentile points of the statistic based on 2000 samples of size n from the generalized Poisson model with parameters α and μ, and the same points of a χ^2_1 distribution.

			Percentile points of a χ^2_1 distribution				
n	α	μ	$p_{.7} = 1.07$	$p_{.8} = 1.64$	$p_{.9} = 2.71$	$p_{.95} = 3.84$	$p_{.99} = 6.63$
100	0.50	0.5	1.10	1.63	2.61	3.73	6.45
	0.80	0.6	1.10	1.64	2.70	3.90	6.19
	0.75	0.8	1.13	1.64	2.72	3.75	6.35
	0.25	1.0	1.05	1.58	2.55	3.90	6.98
	0.25	2.4	1.14	1.72	2.78	3.73	6.36
200	0.50	0.5	1.12	1.73	2.71	3.75	6.48
	0.80	0.6	1.12	1.73	2.74	3.87	6.54
	0.75	0.8	1.06	1.63	2.70	3.80	6.19
	0.25	1.0	1.10	1.60	2.75	3.59	6.10
	0.25	2.4	1.11	1.72	2.88	3.96	6.84

When $\alpha = 0$, the last term in b will be zero as it is obtained from a derivative with respect to α. Also, c will be zero since it is from a derivative with respect to α. Thus, b reduces to $n(e^{\mu} - 1) - n\mu$ and the score statistic in (5.6) reduces to the result obtained by van den Broek (1995) for the Poisson distribution. Under the null hypothesis of generalized Poisson model, the score statistic has an asymptotic chi-square distribution with 1 degree of freedom.

Remark: The information matrix on page 212 of Gupta *et al.* (1996) appears to be incorrect. Only two of the entries in the matrix should be zero as opposed to the four given by them. The authors set $\lambda = (1 - \Phi)(1 - e^{-\theta})$ in the log-likelihood function. However, on differentiating the log-likelihood function with

respect to λ and θ, it appears this functional relationship between λ and θ was not considered.

A simulation study was carried out in order to see if the chi-square approximation is appropriate. From a generalized Poisson distribution (GPD) with the sets of parameter values in Table 2, we generated 2000 samples of sizes $n = 100$ and 200. These sets of parameter values were chosen so that the GPD mean will be low and there will be a lot of zeros in the generated data. For every sample, the score statistic in (5.6) was calculated. The 70-th, 80-th, 90-th, 95-th, and 99-th percentiles are reported in Table 2. These percentile points look reasonable when compared to the percentile points of a chi-square distribution with one degree of freedom. When the mean of the GPD is large, there is hardly any zero. For this situation, the chi-square approximation is not as good. However, there is little or no need for a zero inflation test for large means.

5.2 The Case with covariates

The score function $U(\boldsymbol{\beta}, \alpha, 0)$ and the expected information matrix $I(\boldsymbol{\beta}, \alpha, 0)$ can be calculated from the log-likelihood in (5.1) with replacing μ by $\mu_i = \mu_i(x_i)$, which depends on the covariates. A score test in the inflated generalized Poisson distribution has the advantage that one does not need to fit the ZIGP regression model but just the GPR model which is the distribution under the null hypothesis. The score statistic for testing whether the GPR model fits the number of zeros well is, in this case, given by

$$S_c(\hat{\boldsymbol{\beta}}, \hat{\alpha}) = S_c(\hat{\boldsymbol{\beta}}, \hat{\alpha}, 0) = U'(\hat{\boldsymbol{\beta}}, \hat{\alpha}, 0)[I(\hat{\boldsymbol{\beta}}, \hat{\alpha}, 0)]^{-1})U(\hat{\boldsymbol{\beta}}, \hat{\alpha}, 0), \tag{5.7}$$

where $\hat{\alpha}$ and $\hat{\boldsymbol{\beta}}$ are the maximum likelihood estimates of α and $\boldsymbol{\beta}$ under the null hypothesis of GPR model. Under the null hypothesis, the score statistic has an asymptotic chi-square distribution with 1 degree of freedom.

5.3 Goodness-of-fit test

A measure of goodness-of-fit of the ZIGP regression model may be based on the log-likelihood statistic. The ZIGP regression model in (3.2) reduces to the ZIP regression model when the dispersion parameter $\alpha = 0$. To test for the adequacy of the ZIGP model over the ZIP regression model, one may test the hypothesis $H_0 : \ \alpha = 0$ against $H_a : \ \alpha \neq 0$. The addition of the dispersion parameter α in the regression model is justified if H_0 is rejected. To test the null hypothesis H_0, one can use the likelihood ratio statistic. Alternatively, one can use the asymptotic Wald statistic for parameter α which is calculated after fitting the ZIGP regression model.

Note: The second derivatives of (4.1) and (4.5) with respect to the parameters and the entries of $U(\boldsymbol{\beta}, \alpha, 0)$ and $I(\boldsymbol{\beta}, \alpha, 0)$ in (5.7) are available from the first author.

6. Results and Discussion

The results of using the ZIP and ZIGP models are given in Table 3. The data set has too many zeros (observed proportion of zeros is 66.4%) which led us to apply the ZIP model. The estimated proportions of zeros from ZIP and ZIGP regression models are, respectively, 63.7% and 65.7%. The zero-inflated negative binomial (ZINB) regression model is a competitor to the ZIGP model when there is a situation of over-dispersion and of too many zeros. The domestic violence data are over-dispersed with 66.4% zeros. However, the ZINB regression model did not converge in fitting the data. Lambert (1992) also observed this problem in fitting ZINB regression model to an observed data set. This realization led us to develop and to apply the ZIGP regression model for modeling over-dispersed data with too many zeros.

Table 3: Estimates from ZIP regression and ZIGP regression models

	ZIP		ZIGP	
Variable	Estimate ± SE	t-value	Estimate ± SE	t-value
Intercept	3.4206 ± 0.1729	19.78**	5.4332 ±1.2620	4.31**
Edu_v	-0.3569 ± 0.0550	-6.49**	-1.5005 ± 0.4967	-3.02**
Edu_b	0.0370 ± 0.0527	0.70	0.5907 ± 0.3035	1.95
Emp_v	0.1252 ± 0.0897	1.40	0.3419 ± 0.5027	0.68
Emp_b	0.0211 ± 0.1051	0.20	1.2458 ± 0.7711	1.62
Inc_v	-0.0878 ± 0.0362	-2.43*	-0.4814 ± 0.2154	-2.24*
Inc_b	-0.2012 ± 0.0384	-5.25**	-0.4183 ± 0.2466	-1.70
Fam_v	0.1245 ± 0.0999	1.25	0.1804 ± 0.4629	0.39
Fam_b	-0.1645 ± 0.0696	-2.36*	-0.6656 ± 0.4951	-1.34
Club_v	0.7804 ± 0.1050	7.43**	1.7158 ± 0.7047	2.43*
Club_b	-0.8548 ± 0.1222	-7.00**	-1.9866 ± 0.7128	-2.79**
Drug_v	-0.7577 ± 0.1275	-5.94**	-1.0645 ± 0.5377	-1.98*
Drug_b	0.6305 ± 0.0929	6.79**	1.5428 ± 0.4019	3.84**
τ	-0.2456 ± 0.0619	-3.97**	-0.1242 ± 0.0570	-2.18*
α			0.3050 ± 0.0556	5.49**
Log-likelihood	-641.09		-365.84	

* indicates significant at 0.05 level; ** indicates significant at 0.01 level; SE = standard error

The score statistic in (5.6) is computed from the data and we obtained a value of 20.02. This value is significant at 5% level when compared to the tabulated chi-square distribution with one degree of freedom. By using the score statistic, we conclude that the data have too many zeros and the GPR model is not an appropriate model. Thus, the ZIGP regression model is more appropriate than the GPR model for the domestic violence data. From Table 3, a test of $\alpha = 0$ by using the asymptotic Wald statistic showed that α is significantly different from zero. Based on this test, the ZIP regression model is not an appropriate model for the domestic violence data. The ZIGP regression model fits the data better than the ZIP model with almost one fold increase in the value of the likelihood.

In Table 3, there is a significant negative relationship between the victim's income and the level of violence. Thus, victims with high income tend to receive lower number of violence. This finding is also supported by Farmer and Tiefenthaler (1997). Only the ZIP model, but not the ZIGP model gave a similar conclusion for the batterer's income. The victim's education is negatively related to the level of violence. There is a significant positive relationship between the victim's belonging to a club and the level of violence. However, it is a significant negative relationship for the batterer. In regard to drug problem, there is a significant positive relationship between the batterer having drug problem and the level of violence. This indicates that more drug problems the batterer has, more violent the batterer becomes. The relationship between drug problem and the level of violence is negative and significant for the victim in both regression models. Overall, six independent variables are significant at 1% level under the ZIP model whereas only three are significant at 1% level under the ZIGP model.

7. Conclusion

Even though the ZIGP regression model is a good competitor of ZINB regression model, we do not know under what conditions, if any, which one will be better. The only observation we have in this regard at this time is that in all the datasets fitted to both models, we successfully fitted the ZIGP regression model to all datasets. However, in a few cases, the iterative technique to estimate the parameters of ZINB regression model did not converge. This observation is similar to the one made by Lambert (1992) as we remarked in the introduction section. The application of the ZIGP regression model to the domestic violence data illustrates the usefulness of the model.

Acknowledgment

This work was done while Felix Famoye, Central Michigan University, was on his sabbatical leave at the Department of Biostatistics, School of Public Health,

University of North Texas at Fort Worth.

References

Bohning, D., Dietz, E., Schlattmann, P., Mendonca, L., and Kirchner, U. (1999). The zero-inflated Poisson model and the decayed, missing and filled teeth index in dental epidemiology. *Journal of Royal Statistical Society* A **162**, 195-209.

Consul, P. C. (1989). *Generalized Poisson Distributions: Properties and Applications.* Marcel Dekker.

Consul, P. C. and Famoye, F. (1992). Generalized Poisson regression model. *Communications in Statistics, Theory and Methods* **21**, 89-109.

Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics.* Chapman and Hall.

Famoye, F. (1993). Restricted generalized Poisson regression model. *Communications in Statistics, Theory and Methods* **22**, 1335-1354.

Farmer, A. and Tiefenthaler, J. (1997). An economic analysis of domestic violence. *Review of Social Economy* **55**, 337-358.

Frome, E. L., Kurtner, M. H. and Beauchamp, J. J. (1973). Regression analysis of Poisson-distributed data. *Journal of the American Statistical Association* **68**, 288-298.

Gupta, P. L., Gupta, R. C., and Tripathi, R. C. (1996). Analysis of zero-adjusted count data. *Computational Statistics and Data Analysis* **23**, 207-218.

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

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

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

Heibron, D. (1994). Zero-altered and other regression models for count data with added zeros. *Biometrical Journal* **36**, 531-547.

Hinde, J. P. and Demetrio, C. G. B. (1998). Overdispersion: models and estimation. *Computational Statistics and Data Analysis* **27**, 151-170.

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

Lee, A. H., Wang, K. and Yau, K. K. W. (2001). Analysis of zero-inflated Poisson data incorporating extent of exposure. *Biometrical Journal* **43**, 963-975.

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

Ridout, M., Demetrio, C. G. B. and Hinde, J. (1998). Models for count data with many zeros. Invited paper presented at the Nineteenth International Biometric Conference, Cape Town, South Africa, 179-190.

van den Broek, J. (1995). A score test for zero-inflation in a Poisson distribution. *Biometrics* **51**, 738-743.

Wang, W. and Famoye, F. (1997). Modeling household fertility decisions with generalized Poisson regression. *Journal of Population Economics* **10**, 273-283.

Wulu, J. T., Singh, K. P. Famoye, F. and McGwin, G. (2002). Regression analysis of count data. *Journal of the Indian Society of Agricultural Statistics* **55**, 220-231.

Received Auggust 4, 2004; accepted November 25, 2004.

Felix Famoye
Department of Mathematics
Central Michigan University
Mount Pleasant, MI 48859, USA
fleix.famoye@cmich.edu

Karan P. Singh
School of Public Health
UNT Health Science Center
Fort Worth, TX 76107, USA
ksingh@hsc.unt.edu