

Generalized Poisson-Poisson Mixture Model for Misreported Counts with an Application to Smoking Data

Mavis Pararai¹, Felix Famoye² and Carl Lee²

¹*Indiana University of Pennsylvania* and ²*Central Michigan University*

Abstract: The assumption that is usually made when modeling count data is that the response variable, which is the count, is correctly reported. Some counts might be over- or under-reported. We derive the Generalized Poisson-Poisson mixture regression (GPPMR) model that can handle accurate, underreported and overreported counts. The parameters in the model will be estimated via the maximum likelihood method. We apply the GPPMR model to a real-life data set.

Key words: Generalized Poisson regression, regression, underreporting.

1. Introduction

Many real world applications involve count data. There are a lot of regression models that have been used in modeling count data. Some of these regression models have been applied to data on number of bottles of port wine purchased (Ramos, 1999); number of absenteeism in the workplace (Barmby et al, 1991); underreporting of needlestick injuries by medical students (Watermann, Jankowski and Madan, 1994), the frequency of criminal victimization, (Li, Trivedi and Guo, 2003), to mention only a few. A good compilation on regression analysis of count data is given by Cameron and Trivedi, (1998). In most of these cases, the number of counts could have been potentially overreported, underreported or correctly reported. In the case of the counts having been correctly reported, then the appropriate count data regression model such as negative binomial, Poisson and generalized Poisson can be applied to such data. In real life there is potential of misreporting and it is necessary to check count data for this kind of reporting.

Winkelmann (1996) proposed a Poisson regression model that takes underreporting into account. This model is a mixture of the Poisson and the binomial distributions. The number of reported events, y_i , that result only if absenteeism occurs was assumed to be Poisson distributed with probability π_i , captured by

the binomial distribution that each individual event is reported. The Poisson regression model for underreported counts is given by

$$P(Y = y_i) = \frac{e^{\pi_i \mu_i} (\pi_i \mu_i)^{y_i}}{y_i!} \quad \text{for } y_i = 0, 1, 2, \dots, \quad (1.1)$$

with mean, $E(Y_i) = \pi_i \mu_i$, where $\mu_i = \mu_i(x_i) = \exp(\sum_{j=1}^k x_{ij} \beta_j)$ and $\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \sum_{j=1}^m z_{ij} \delta_j$.

The negative binomial regression model that takes underreporting into account (Mukhopadhyay, 1997) was derived as a mixture of the negative binomial and the binomial distributions. The resulting mixture regression model for underreported counts is the negative binomial regression model given by Mukhopadhyay (1997) as

$$P(Y = y_i) = \frac{\Gamma(y_i + \alpha^{-1})}{\Gamma(y_i + 1)\Gamma(\alpha^{-1})} \left(\frac{\alpha^{-1}}{\alpha^{-1} + \pi_i \mu_i}\right)^{\alpha^{-1}} \left(\frac{\pi_i \mu_i}{\alpha^{-1} + \mu_i \pi_i}\right)^{y_i} \quad \text{for } y_i \geq 0. \quad (1.2)$$

where α is the dispersion parameter and $0 < \pi_i < 1$ is the probability of underreporting an event and is conditional on some covariates $z_i = (z_{i1}, z_{i2}, \dots, z_{im})$ and $\mu_i = \mu_i(x_i) = \exp(\sum_{j=1}^k x_{ij} \beta_j)$. The probability π_i is modeled through the logit link function specification. The mean and variance of this model are given by Mukhopadhyay (1997) as $E(Y_i) = \pi_i \mu_i$ and $Var(Y_i|x_i, z_i) = \pi_i \mu_i (1 + \alpha \pi_i \mu_i)$. Mukhopadhyay (1997) applied this regression model to a data set from the National Longitudinal Survey for Youth for the year 1980 with the response variable being the number of times one had been convicted of some illegal activity.

The generalized Poisson regression (GPR) model (Famoye, 1993) is given by

$$P(Y = y_i) = \left(\frac{\mu_i}{1 + \alpha \mu_i}\right)^{y_i} \frac{(1 + \alpha y_i)^{y_i - 1}}{y_i!} \exp\left[\frac{-\mu_i(1 + \alpha y_i)}{1 + \alpha \mu_i}\right], \quad (1.3)$$

for $y_i \geq 0$ and μ_i is the log-link function. When $\alpha = 0$, the GPR model becomes the Poisson regression model. When $\alpha > 0$ the GPR model can be used for overdispersed data and when $\alpha < 0$, the GPR model can be used for underdispersed data. The generalized Poisson regression model for underreported counts (GPRUC model) was derived by Pararai, Famoye and Lee (2006). The GPRUC model was applied to data on number of sexual partners. The models mentioned are appropriate if the counts are underreported.

Li, Trivedi and Guo (2003) suggested a mixture model of the Poisson and negative binomial regression models that can be used to handle data that is under-, over- and accurately reported. In the regression model by Li *et al.* (2003), misreporting would occur when an individual reports the number of events

as y_i , $i = 1, 2, \dots, n$ which may differ from the true count y_i^* , $i = 1, 2, \dots, n$. The negative binomial regression model took care of the accurate counts while the Poisson regression model took care of the underreported and overreported counts. The means of the accurate, overreported and underreported counts were given respectively by $\lambda_i = \exp(x_{ij}\gamma_j)$, $\mu_i = \exp(z_{ij}\delta_j)$ and $\psi_i = y_i^* \exp(x_{ij}\beta_j)$, where x_{ij} and z_{ij} represent the covariates on which these means depend. The regression model for handling data with accurately reported, overreported and underreported counts derived by Li *et al.* (2003) is given as

$$\begin{aligned}
 P(Y = y_i | x_i, z_i, \gamma, \delta, \beta, \alpha) &= \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!} \left(\frac{\alpha^{-1}}{\alpha^{-1} + \lambda_i} \right)^{\alpha^{-1}} + \sum_{y_i^*=1}^{\infty} \left[\frac{e^{-y_i^*} \xi_i \psi_i^{y_i}}{y_i^*!} \right. \\
 &\times \left. \frac{\Gamma(y_i^* + \alpha^{-1})}{\Gamma(y_i^* + 1)\Gamma(\alpha^{-1})} \left(\frac{\alpha^{-1}}{\alpha^{-1} + \lambda_i} \right)^{\alpha^{-1}} \left(\frac{\lambda_i}{\alpha^{-1} + \lambda_i} \right)^{y_i} \right].
 \end{aligned}
 \tag{1.4}$$

Li *et al.* (2003) applied the regression model in (3.1) to school crime victimization data drawn from the National Crime Victimization Survey for the year 1995. The response variable was the number of stolen items from one's locker in school.

In many cases the negative binomial regression model and the generalized Poisson regression model are competitors when fitting count data. It is therefore reasonable to derive a generalized Poisson regression model that accommodates misreported counts along the same way as its negative binomial counterpart by Li, Trivedi and Tong (2003).

The remainder of the paper is organized as follows: In section 2 a description of the National Pregnancy and Health Survey (NPHS) data is given. Section 3 gives an outline of how the GPPMR model is developed. The parameters of the model are estimated via the maximum likelihood method and this is explained in section 4. Some goodness-of-fit tests are given in section 5. The GPPMR model is applied to the NPHS data in section 6 and the results are also discussed. The concluding remarks will be given in section 7.

2. National Pregnancy and Health Survey Data

The data was collected from the National Pregnancy and Health Survey: Drug Use Among Women Delivering Live Births, 1992. The data can be accessed from <http://webapp.icpsr.umich.edu/cocoon/icpsr-study/02835.xml>. One of the objectives of the study was to describe the use of illegal drugs by expecting mothers. The data on substance use was collected through a questionnaire that was administered to women during pregnancy. One of the variables measured

Table 1: Description of variables for the NPHS data

Variable	Definition
numcig (y)	Number of cigarettes smoked a day
mstatus	Marital status: 1=married, 0=unmarried
hispanic	Race: 1=Hispanic, 0=Other
black	Race: 1=Black, 0=Other
college	Has a college degree: 1=yes, 0=no
wages	Source of income was wages and salaries, 1=yes, 0=no
public	Received public assistance or welfare, 1=yes, 0=no
foodstamp	Received food stamps, 1=yes, 0=no
housing	Received housing assistance, 1=yes, 0=no
ssi	Received supplementary income, 1=yes, 0=no
unemp	Received unemployment insurance, 1=yes, 0=no
livesmoker	Lived with a smoker, 1=yes, 0=no
help	Tried to get help to quit smoking, 1=yes, 0=no
last3smoke	Smoked in the last 3 months of pregnancy, 1=yes, 0=no

Table 2: Descriptive statistics for the NPHS data

Variable	Mean	Std Deviation	% of 1's
numcig(y)	4.2895	9.2456	
smoke			23.58
mstatus			66.57
hispanic			16.40
black			19.10
college			19.57
wages			91.08
public			17.16
foodstamp			22.74
housing			4.78
ssi			6.26
unempinc			7.14
livesmoker			33.39
help			1.23
last3smoke			14.29

was the number of cigarettes a woman smoked each day during the first trimester of pregnancy. To demonstrate the GPPMR model, the NPHS data is considered with the number of cigarettes a woman smoked in the first trimester of pregnancy as the response variable. The explanatory variables and the response variable used in modeling the data are described in Table 1. The descriptive statistics for this data are shown in Table 2.

The mean of the number of cigarettes smoked in Table 2 is less than its variance showing that the data is overdispersed. The variables chosen in illustrating

the GPPMR model pertain to source of income of the respondent.

3. Generalized Poisson-Poisson Mixture Model

The generalized Poisson-Poisson mixture regression (GPPMR) model accommodating over-, under- and accurately reported counts is a mixture of the generalized Poisson regression model in (1.3) and the Poisson regression model. The justification for mixing Poisson and generalized Poisson is that we want two data generating processes that result in count data. The Poisson model provided the most reasonable choice after trying other models such as negative binomial and generalized Poisson. Also, in the simulations that were carried out, convergence was much quicker when mixing the generalized Poisson and Poisson models. The assumptions used in deriving the GPPMR model are the same as those used by Li *et al.* (2003) in deriving the NBPMM model in (3.1). Let y_i^* denote the total number of true events for individual i where $i = 1, 2, \dots, n$. Assume that y_i^* conditional on covariates $x_i = (x_{i1}, x_{i2}, \dots, x_{ik})$ follows the generalized Poisson distribution with probability function

$$P(y_i^* | x_i) = \left(\frac{\lambda_i}{1 + \alpha \lambda_i} \right)^{y_i^*} \frac{(1 + \alpha y_i^*)^{y_i^* - 1}}{y_i^*!} \exp \left[\frac{-\lambda_i (1 + \alpha y_i^*)}{1 + \alpha \lambda_i} \right], \quad (3.1)$$

where the mean function $\lambda_i = \exp(x_i \gamma)$ and γ is a k -dimensional vector of unknown regression coefficients. The variance of the regression model in (6) is $\lambda_i (1 + \lambda_i)^2$. The distribution of y_i^* in (3.2) will be denoted as $GPR[\lambda_i, \lambda_i (1 + \lambda_i)^2]$. The counts are reported incorrectly when an individual reports the number of an event as y_i , different from y_i^* , $i = 1, 2, \dots, n$.

Assume that when $y_i^* = 0$, the observed count y_i is Poisson distributed with mean and variance $\mu_i = \exp(z_i \delta)$ denoted by $P(\mu_i)$ where $z_i = (z_{i1}, \dots, z_{ip})$ are some explanatory variables and δ is a vector of some unknown parameters. This is a situation when potential overreporting may occur since an individual is reporting a value y_i while the true value is y_i^* . Furthermore, conditional on $y_i^* > 0$, y_i is Poisson distributed with mean and variance given by $y_i^* \exp(z_i \beta)$ (Li, Trivedi and Guo, 2003). This distribution shall be denoted by $P(y_i^* \xi_i)$ where $\xi_i = \exp(z_i \beta)$ is dependent on the covariates $z_i = (z_{i1}, \dots, z_{ip})$ and β is a vector of unknown parameters. This is a situation where potential underreporting of events occurs. The covariates used in modeling the accurate portion of the regression model maybe the same as those used in modeling the over and underreported portions. These assumptions from Li *et al.* (2003) can be summarized as:

- (1) Overreporting occurs for $y_i | y_i^* = 0 \sim P[\exp(z_i \delta)] = P(\mu_i)$,
- (2) Underreporting occurs for $y_i | y_i^* > 0 \sim P[y_i^* \exp(z_i \beta)] = P(y_i^* \xi_i)$,
- (3) Accurate reporting occurs for $y_i^* \sim GPR[\lambda_i, \lambda_i (1 + \lambda_i)^2]$.

The probability distribution of the reported count y_i can be obtained as the marginal density of the joint distribution of the generalized Poisson and the Poisson regression models. The model for the reported counts is:

$$\begin{aligned}
 P(y_i|x_i, z_i, \gamma, \delta, \beta, \alpha) &= \sum_{y_i^*=0}^{\infty} P(y_i|y_i^*, z_i, \delta, \beta)P(y_i^*|x_i, \gamma, \alpha) \\
 &= P(y_i|y_i^*, z_i, \delta, \beta)P(y_i^* = 0|x_i, \gamma) \\
 &+ \sum_{y_i^*=1}^{\infty} P(y_i|y_i^*, z_i, \delta, \beta)P(y_i^*|x_i, \gamma, \alpha) \\
 &= \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!} \left(\frac{-\lambda_i}{1 + \alpha\lambda_i} \right) + \sum_{y_i^*=1}^{\infty} \left[\frac{(y_i^* \xi_i)^{y_i} e^{-(y_i^* \xi_i)}}{y_i!} \right. \\
 &\quad \left. \times \left(\frac{\lambda_i}{1 + \alpha\lambda_i} \right)^{y_i^*} \frac{(1 + \alpha y_i^*)^{y_i^*-1}}{y_i^*!} \exp \left[\frac{-\lambda_i(1 + \alpha y_i^*)}{1 + \alpha\lambda_i} \right] \right\} \quad (3.2)
 \end{aligned}$$

The mean and variance of the GPPMR model are

$$\begin{aligned}
 E(Y_i|x_i, z_i) &= E(Y_i|Y_i^* = 0) + E(Y_i|Y_i^* > 0) \\
 &= \exp \left(\frac{-\lambda_i}{1 + \alpha\lambda_i} \right) \mu_i + \xi_i \lambda_i. \quad (3.3)
 \end{aligned}$$

and

$$\begin{aligned}
 Var(Y_i|x_i, z_i) &= E(Y_i^2|x_i, z_i) - [E(Y_i|x_i, z_i)]^2 \\
 &= \exp \left(\frac{-\lambda_i}{1 + \alpha\lambda_i} \right) \mu_i + \xi_i \lambda_i + \mu_i^2 \exp \left(\frac{-\lambda_i}{1 + \alpha\lambda_i} \right) \\
 &\quad \times \left[1 - \left(\frac{-\lambda_i}{1 + \alpha\lambda_i} \right) \right] \\
 &+ \xi_i \lambda_i \left[\xi_i (1 + \alpha\lambda_i)^2 - 2\mu_i \exp \left(\frac{-\lambda_i}{1 + \alpha\lambda_i} \right) \right], \quad (3.4)
 \end{aligned}$$

respectively.

4. Estimation of Model Parameters

The log-likelihood function of the GPPMR model in (3.3) is

$$\begin{aligned}
 L(y_i, \gamma, \delta, \beta, \alpha) &= \sum_{i=0}^n \log \left\{ \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!} \left(\frac{-\lambda_i}{1 + \alpha\lambda_i} \right) + \sum_{y_i^*=1}^{\infty} \left[\frac{(y_i^* \xi_i)^{y_i} e^{-(y_i^* \xi_i)}}{y_i!} \right. \right. \\
 &\quad \left. \left. \times \left(\frac{\lambda_i}{1 + \alpha\lambda_i} \right)^{y_i^*} \frac{(1 + \alpha y_i^*)^{y_i^*-1}}{y_i^*!} \exp \left(\frac{-\lambda_i(1 + \alpha y_i^*)}{1 + \alpha\lambda_i} \right) \right] \right\} \quad (4.1)
 \end{aligned}$$

To estimate the parameters α, β, δ and γ , the Statistical Analysis Software (SAS, 1999) was used. The NLPNRA algorithm, which is a nonlinear optimization based on the Newton-Raphson method is used to estimate the parameters α, β, δ and γ . The variance-covariance matrix of the estimated parameters is obtained from the NLPFDD subroutine in SAS. This subroutine can approximate derivatives by using finite differences and computes the gradient vector and the Hessian matrix H , all evaluated at $\hat{\alpha}, \hat{\beta}, \hat{\delta}$ and $\hat{\gamma}$.

5. Goodness-of-fit Tests

The GPPMR model in (3.3) reduces to the Poisson-Poisson mixture regression model when the dispersion parameter $\alpha = 0$. To assess the appropriateness of the GPPMR model over the Poisson-Poisson mixture regression model one can test the hypothesis: $H_0 : \alpha = 0$ against $H_a : \alpha \neq 0$. To carry out the test, one fits the GPPMR model and uses the asymptotic Wald t-test. The statistic to be computed is given by

$$W = \frac{\hat{\alpha}}{se(\hat{\alpha})}, \quad (5.1)$$

where $\hat{\alpha}$ is the maximum likelihood estimate of α and $se(\hat{\alpha})$ is its corresponding standard error. This statistic is compared to the t distribution with $n - \nu - 1$ degrees of freedom, where ν is the total number of parameters in the GPPMR model.

The GPPMR and NBPMR models for underreported, overreported and accurate counts are non-nested. In order to discriminate between the two non-nested models the Vuong (1989) test will be used. The hypothesis to be tested is H_0 : GPPMR and NBPMR models are equivalent against the two alternatives H_f : GPPMR model is better than NBPMR model, or H_g : NBPMR model is better than GPPMR model, where H_f and H_g are the two competing alternative hypotheses for the GPPMR and NBPMR models respectively.

6. Results and Discussion

The independent variables described in Table 1 are taken as the covariates that affect λ_i , the mean of the accurately reported counts. The covariates marital status, hispanic, black, wages and salaries, food stamps, and unemployment income are used to model μ_i and $y_i^* \exp(z_i \beta)$ which represent the mean of the overreported and underreported counts respectively. The proportion of zeros in the response variable was 76.42%. The results obtained from fitting the GPPMR and NBPRM models are shown in Table 3.

Table 3: Estimates for NBPRM and GPPRM models

Variables	NBPRM Model Estimate \pm Std Error	GPPRM Model Estimate \pm Std Error
Accurate Reporting		
const	0.4859 \pm 0.1601*	0.5234 \pm 0.1570*
mstatus	-0.5712 \pm 0.1398*	-0.5686 \pm 0.1233*
hispanic	-1.1976 \pm 0.1750*	-1.1916 \pm 0.1688*
black	-0.9493 \pm 0.1453*	-0.9837 \pm 0.1383*
college	-0.6263 \pm 0.1616*	-0.5968 \pm 0.1564*
wages	-1.8169 \pm 0.1678*	-1.8861 \pm 0.1540*
public	0.0305 \pm 0.1429	0.0366 \pm 0.1413
foodstamps	0.4007 \pm 0.1600*	0.3094 \pm 0.1413*
housing	0.2640 \pm 0.1523	0.2752 \pm 0.1557
ssi	-0.1810 \pm 0.1493	-0.2083 \pm 0.1469
unempinc	0.2631 \pm 0.1669	0.2733 \pm 0.1627
livesmoker	0.8180 \pm 0.093*	0.8040 \pm 0.0868*
help	0.6253 \pm 0.2504*	0.6163 \pm 0.2489*
last3smoke	2.0865 \pm 0.0865*	2.1055 \pm 0.0835*
Overreporting		
mstatus	-1.3579 \pm 0.7357	-1.3675 \pm 0.7747
hispanic	-1.0944 \pm 0.7394	-1.0842 \pm 0.7406
black	-1.4933 \pm 0.7757	-1.4460 \pm 0.7580
wages	-4.1371 \pm 0.6360*	-4.1361 \pm 0.6309*
foodstamps	15.8541 \pm 152.1260	-14.3106 \pm 138.2111
unempinc	0.3289 \pm 1.0975	0.3686 \pm 1.0996
Underreporting		
mstatus	0.2244 \pm 0.0978*	0.1949 \pm 0.0786*
hispanic	0.0662 \pm 0.0913	0.0427 \pm 0.0839
black	0.0025 \pm 0.1007	0.0172 \pm 0.0928
wages	1.9564 \pm 0.1035*	2.0100 \pm 0.0816*
foodstamps	-0.1934 \pm 0.0917*	-0.1233 \pm 0.0825
unempinc	0.0018 \pm 0.0964	-0.0010 \pm 0.0908
α	1.3092 \pm 0.1956	0.1930 \pm 0.0282*
Log Likelihood	-3116.535	-3128.560

A test of the null hypothesis $\alpha = 0$ by using the asymptotic Wald t - statistic shows that α , the dispersion parameter, is different from zero in Table 3. The Poisson-Poisson Mixture Regression model is not appropriate and hence cannot be used to describe this data based on the Wald's t - test result. A comparison between the fitted NBPMR and the GPPMR models is made using the Vuong (1989) test. The Vuong (1989) statistic calculated is equal to -0.0363. Since $|T_*| = 0.0363 < Z_{0.025} = 1.96$, the null hypothesis that states that the GPPMR and the NBPMR models are equivalent cannot be rejected. This result shows that

the NBPMR and GPPMR models are similar in their performance. In Table 3 the log-likelihood values for the NBPMR and GPPMR models are given respectively by -3116.535 and -3128.56 showing no significant difference in the performance of the two models.

6.1 Results on accurate reporting

The probability of accurately reporting the number of cigarettes smoked in a day by a woman during the first 3 months of pregnancy is given by $P(y_i^*|x_i, \gamma)$. The results from the GPPMR model in Table 3 suggest that this probability is greater among women who lived with a smoker, tried to get help to quit smoking, smoked in the last 3 months of pregnancy, received food stamps, unmarried women, non-hispanics, non-blacks, women with no college education and women who do not have wages and salaries as sources of income. This probability does not seem to be affected by women who receive social security income, public assistance, housing assistance and unemployment income.

6.2 Results on overreporting and underreporting

The probability of overreporting of events is given by $P(y_i|y_i^* = 0, z_i, \delta)$. The GPPMR model shows the only covariate that affects the probability of overreporting the number of cigarettes smoked by a pregnant woman in the first 3 months of pregnancy is wages and salaries. The probability of overreporting an event is negatively related to wages and salaries. Women who did not have wages and salaries as sources of income tend to overreport the number of cigarettes smoked in a day during the first 3 months of pregnancy.

The probability of underreporting the number of cigarettes smoked is positively related to marital status and wages and salaries. Married women who smoked during the first 3 months of pregnancy tend to underreport the number of cigarettes smoked in a day compared to their unmarried counterparts.

7. Concluding Remarks

In this paper we presented and examined the mixture model of the Poisson and generalized Poisson regression models, namely, the generalized Poisson-Poisson mixture regression model. This was an attempt to come up with a model that can be used to model counts that could be potentially misreported. The model is capable of capturing all 3 potential forms of reporting namely, accurate, under- and over-reporting. Other methods of estimating the parameters other than the maximum likelihood method could also be explored. The issue of variable selection could further be explored in as far as determining how to choose the variables

that affect the accurate, under- and over-reported portions of the model. Other data sets could possibly yield results in which the GPPMR model outperforms the NBPMR model.

Acknowledgement

The authors express their gratitude to the editor and referee for many useful comments and suggestions.

References

- Barmby, T.A., Orme, C., and Treble, J.G. (1991). Worker absenteeism: An analysis using microdata. *Economic Journal*, **101**, 214-229.
- Cameron, A.C., and Trivedi, P.K. (1998). *Regression analysis of count data.*, Cambridge, UK: Cambridge University Press.
- Famoye, F. (1993). Restricted generalized Poisson regression model. *Communications in Statistics-Theory and Methods*, **22**, 1335-1354.
- Li, T., Trivedi, P.K., and Guo, J. (2003). Modeling response bias in count: A structural approach with an application to the national crime victimization survey data. *Sociological Methods and Research* **31**, 514-544.
- Mukhopadhyay, K.(1997). *Regression models for underreported Counts*. Unpublished doctoral dissertation, Indiana University, Indiana.
- Pararai, M., Famoye, F. and Lee, C. (2006). Generalized Poisson regression model for underreported counts. *Advances and Applications in Statistics* **6**, 305-322.
- Ramos, F.F.R. (1999). Underreporting of purchases of port wine. *Journal of Applied Statistics* **26**, 485-494.
- SAS institute Inc., *SAS/IML User's Guide, Version 8*, Cary, NC: SAS Institute Inc., 1999. 846 pp.
- Vuong, Q.H. (1989). Likelihood ratio tests for model selection and non-nested hypothesis. *Econometrica* **57**, 307-333.
- Watermann, J., Jankowski, R., and Madan, I. (1994). Underreporting of needlestick injuries by medical students. *Journal of Hospital Infection* **26**, 149-151.
- Winkelmann, R. (1996). Markov chain Monte Carlo analysis of underreported count data with an application to worker absenteeism. *Empirical Economics* **21**, 575-587.

Received September 15, 2008; accepted June 14, 2009.

Mavis Pararai
Department of Mathematics
Indiana University of Pennsylvania Indiana, PA 15705, USA
pararaim@iup.edu

Felix Famoye
Department of Mathematics
Central Michigan University
Mount Pleasant, MI 48859, USA
famoy1kf@cmich.edu

Carl Lee
Department of Mathematics
Central Michigan University
Mount Pleasant, MI 48859, USA
Lee1c@cmich.edu