On the generalized Poisson regression model

Felix Famoye
Department of Mathematics, Central Michigan University, Mount Pleasant, Michigan 48859

Abstract

The paper discusses some of the applications of the generalized Poisson regression model and its various forms of modifications to grouped data, censored data, truncated data and inflated data. Some tests to discriminate between the generalized Poisson regression model and its competitors are reviewed.

Keywords: *estimation*; *tests*; *inflated data*; *applications.*

1. Introduction

In count data regression analysis, the dependent variable is a count where we assume a discrete distribution error structure. The independent variables could take any form (discrete or continuous). In the linear regression analysis, we assume that the error terms have normal distribution. Here, this assumption is not made since the response variable is nonnegative.

The Poisson distribution has been applied in the context of regression analysis for describing count data where the sample mean and sample variance are almost equal [See Cameron and Trivedi (1998) and the references there in.] In many situations, count data are over-dispersed or under-dispersed. Over-dispersion relative to the Poisson is when the sample variance is substantially in excess of the sample mean. Under-dispersion is a situation in which the sample variance is less than the sample mean. Many regression models have been suggested to deal with over-dispersion or under-dispersion. Among these various models are the negative binomial regression model defined and studied by Lawless (1987) and the generalized Poisson regression model studied by Famoye (1993).

Suppose that $y_1, y_2, \ldots, y_n$ are the independent responses, with discrete distribution. A general count regression model can be written as $E(Y_i) = \mu(x_i) = c_i f(x_i, \beta)$, i = 1, 2, …, n, where c_i is a measure of exposure, $x_i = (1, x_1, x_2, \ldots, x_{p-1})$ is a px1 vector of independent variables, $f(x_i, \beta)$ is a differentiable function of p-dimensional vector $\beta = (\beta_0, \beta_1, \beta_2, \ldots, \beta_{p-1})$ of regression parameters. For count regression model, we can write $\mu(x_i) = e^{X_i'\beta}$. Hence, taking natural logarithm, we have

$$\log[\mu(x_i)] = X_i'\beta\,, \tag{1}$$

where the β_j, j = 0, 1, 2, …, p – 1 is the parameter which represents the expected change in the log of the mean per unit increase (or decrease) in the explanatory variable x_j. The model in (1) is often referred to as a count regression model. It is a generalized linear model with link function $g(x)$, where, $g(x) = \log(x)$.

In this paper, we will briefly mention the Poisson regression model, discuss the generalized Poisson regression and some of its modifications to model count data. Finally, we mention two comparison tests for non-nested models.

2. Poisson regression model

The Poisson distribution is perhaps the most used discrete distribution because of its simplicity. The Poisson probability mass function is given by

$$P(Y = y) = \theta^y e^{-\theta} / y!, y = 0, 1, 2, \ldots \qquad (2)$$

The mean and variance of the model in (2) are equal to θ.

Thus, for Poisson regression, we have

$$E(Y_i) = \mu(x_i) = c_i f(x_i, \beta) = e^{x'\beta} = e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_{p-1} x_{p-1}}.$$

Hence, the Poisson regression model is

$$P(Y = y_i \mid x_i) = [\mu(x_i)]^{y_i} e^{-\mu(x_i)} / y_i! \text{ for } y_i = 0, 1, 2, \ldots \qquad (3)$$

The mean and variance of the Poisson regression model in (3) are equal to $\mu(x_i)$. The parameters β can be estimated by the method of maximum likelihood estimation.

Limitation:

The Poisson regression model is adequate for data where mean and variance are about equal. The Poisson distribution is said to be equi-dispersed since the mean and variance are equal. Suppose this is not the case, one needs to use a different model that can handle the type of dispersion.

Over-dispersion in Poisson regression:

When over-dispersion occurs, the standard errors of the parameter estimates are often underestimated and this often leads to wrong conclusions with regards to the significance of the predictor variables. Over-dispersion often arises when modeling count data. Suppose the outcome Y_i has a Poisson distribution with mean μ. Then, V(Y_i) = E(Y_i) = μ. Suppose however, that the variance is proportional to the mean, say, V(Y_i) = φ E(Y_i) = $\varphi\mu$; where φ is to be estimated from our data. There is no dispersion when φ = 1. We observe that over-dispersion occurs when φ > 1, while under-dispersion occurs when 0 < φ < 1.

What may give rise to over-dispersion?

Over-dispersion in Poisson regression could arise as a result of several reasons or combinations of reasons. Over-dispersion in Poisson regression may be caused by positive correlation between the count responses or large variations between count responses. Hilbe (2007) addressed real over-dispersion and apparent over-dispersion. Among the causes of over-dispersion, Hilbe (2007) stated the following:

Real over-dispersion-

- Positive correlation between the counts or excess variation between the counts.
- Violations in the distributional assumptions of the count data.

Apparent over-dispersion-

- The model omits important explanatory variables.
- The data include outliers.
- The model fails to include a sufficient number of interaction terms.
- A predictor variable needs to be transformed.
- The link function is miss-specified.

3. Generalized Poisson regression model

The probability mass function for the generalized Poisson distribution (GPD) is given by

$$P(Y=y)=\theta^{y}(1+\alpha y)^{y-1}e^{-\theta(1+\alpha y)}/y!, y = 0, 1, 2, \ldots \quad (4)$$

The GPD in (4) was defined and studied by Consul (1989). The parameter θ > 0 and the parameter α can be negative or positive. The mean and variance of the GPD in (4) are respectively given by $\mu=\theta/(1-\alpha\theta)$ and $\sigma^{2}=\theta/(1-\alpha\theta)^{3}$. Suppose the mean of the GPD depends on some independent variables x_i, then we can write

$$\mu(x_i)=\theta/(1-a\theta)\Rightarrow\theta=\mu(x_i)/[1+a\mu(x_i)].$$

Thus the generalized Poisson regression (GPR) model can be written as

$$P(Y=y_i\,|\,x_i)=f(y_i)=\left(\frac{\mu(x_i)}{1+\alpha\mu(x_i)}\right)^{y_i}\frac{(1+\alpha y_i)^{y_i-1}}{y_i!}\exp\left[\frac{-\mu(x_i)(1+\alpha y_i)}{1+\alpha\mu(x_i)}\right],$$

$$y_i = 0, 1, 2 \ldots \quad (5)$$

The mean and variance of the generalized Poisson regression model are respectively given by $\mathrm{E}(Y)=\mu(x_i)$ and $\mathrm{V}(Y)=\mu(x_i)[1+\alpha\mu(x_i)]^{2}$. The GPR model reduces to the Poisson regression model when α = 0 and the dispersion factor is given by V(Y_i)/E(Y_i) = $(1+\alpha\mu_i)^{2}$. If α > 0, then V(Y_i) > E(Y_i) and the GPR will model count data with over-dispersion. Similarly, when α < 0, then V(Y_i) < E(Y_i) and the GPR will in this case model under-dispersed count data.

Famoye (1993) applied the GPR in (5) to model the number of faults in rolls of fabric, an over-dispersed count data. Wang & Famoye (1997) considered the GPR model for under-dispersed data on household fertility decisions. The response variable in this application is the number of children up to age 17 years old in a family.

4. Modified generalized Poisson regression model

Examples of data for which modified GPR model can be applied include inflated count data, truncated count data, censored count data, and grouped or categorized count data.

Inflated generalized Poisson regression model:

Over-dispersion in a Poisson model can also arise as a result of too many occurrences of zeros than would normally be expected from a Poisson model. That is, there could be too many zeros than can be assumed theoretically or expected under such model. If this happens, we would then say that the Poisson model is in this case, zero inflated. In some cases, these zeros can be structural zeros in which it is impossible to observe an occurrence. For instance, in a survey to determine the number of bottles of alcohol consumed by respondents per week, there would be individuals in the sample who do not drink alcohol at all. Such people will have the number recorded as zeros but in actual fact, such zeros will be structural as we naturally do not expect them to have a count. On the other hand, if an individual does drink alcohol but he/she did not drink a single bottle of alcohol during the survey period, then such individuals would have a count of zero and the zeros in this case would be referred to as sampling zeros.

In general, the inflation could occur at any point *y* = *k*. Famoye and Singh (2003) proposed a *k*-inflated generalized Poisson regression (*k*-IGPR) to model count data with too many *k*-values. A score test is presented to test whether the number of *k*-values is too large for the generalized Poisson regression model to adequately fit the data. The *k*-inflated generalized Poisson regression model is illustrated using a dataset with too many ones.

A *k*-inflated generalized Poisson regression (*k*-IGPR) model is defined as

$$P(Y = y_i \mid x_i, z_i) = \begin{cases} \varphi_i + (1-\varphi_i)\, f(k;\mu_i,\alpha), & y_i = k \\ (1-\varphi_i)\, f(y_i;\mu_i,\alpha), & y_i \neq k, \end{cases} \tag{6}$$

where $f(y_i;\mu_i,\alpha) = f(y_i)$, y_i = 0, 1, 2, … is the GPR model in (5) and 0 < φ_i < 1. The model in (6) will allow for a decreasing proportion of *k*-values if $-f(k;\mu_i,\alpha)[1- f(k;\mu_i,\alpha)]^{-1}$ < φ_i < 1. In model (6), the functions $\mu_i = \mu_i(x_i)$ and $\varphi_i = \varphi_i(z_i)$ satisfy $\log(\mu_i) = \Sigma_{j=1}^{k} x_{ij}\beta_j$ and $\text{logit}(\varphi_i) = \log\left(\varphi_i / (1-\varphi_i)\right) \; = \Sigma_{j=1}^{m} z_{ij}\delta_j$ where $z_i = (z_{i1} = 1, z_{i2}, z_{i3}, ..., z_{im})$ is the i^{th} row of covariate matrix $\boldsymbol{Z}$ and $\boldsymbol{\delta} = (\delta_1, \delta_2, ..., \delta_m)$ are unknown *m*-dimensional vector of parameters. In this set up, the functions φ_i and μ_i are, respectively, modeled via logit and log link functions. Both are linear functions of some covariates.

Suppose we have $\text{E}(Y_i) = \mu_*$ and $\text{V}(Y_i) = \sigma_*^2$ for the non-inflated model, then the mean and variance of the *k*-inflated GPR model can be written as $\text{E}(Y_i) = \varphi_i k + (1-\varphi_i)\mu_*$ and $\text{V}(Y_i) = \varphi_i k^2 + (1-\varphi_i)(\mu_*^2 + \sigma_*^2) - [\varphi_i k + (1-\varphi_i)\mu_*]^2$ respectively. The mean and variance of the *k*-IGPR model in (6) are given, respectively, by

$$\text{E}(Y_i \mid x_i) = \varphi_i k + (1-\varphi_i)\mu_i(x_i)\,, \tag{7}$$

and

$$\begin{aligned} \text{V}(Y_i \mid x_i) &= \varphi_i k^2 + (1-\varphi_i)\left[\mu_i^2 + \mu_i(1+\alpha\mu_i)^2\right] - \left[\varphi_i k + (1-\varphi_i)\mu_i\right]^2 \\ &= \varphi_i(1-\varphi_i)(k-\mu_i)^2 + (1-\varphi_i)\mu_i(1+\alpha\mu_i)^2\,. \end{aligned} \tag{8}$$

When *k* = 0, the above results for the *k*-inflated GPR model reduce to those given by Famoye and Singh (2006) for the zero-inflated generalized Poisson regression model. The *k*-inflated GPR model reduces to the GPR model when φ_i = 0. It reduces to the *k*-inflated Poisson regression model when α = 0. For positive values of φ_i, it represents the *k*-inflated generalized Poisson regression model and for negative values of φ_i , it represents *k*-deflated generalized Poisson regression model. The *k*-deflation cases rarely occur in practice.

The covariates affecting φ_i and μ_i may or may not be the same. If one does not know the covariates that may affect φ_i one can fit a constant function with $\text{logit}(\varphi_i) = z_{i1}\delta_1 = \delta_1$. If y_i are independent random variables having a *k*-inflated generalized Poisson distribution, the *k*-values are assumed to occur in two distinct states. The only occurrences in the first state are *k*-values which occur with probability φ_i. These can be referred to as 'structural' *k*-values. The second state occurs with probability $(1-\varphi_i)$ and leads to a generalized Poisson distribution with parameters α and μ_i. The *k*-

values from the second state, i.e. from the generalized Poisson distribution, can be called 'sampling' *k*-values. The two-state process leads to a two-component mixture distribution with probability mass function given in (6).

The zero-inflated generalized Poisson regression model is defined (Famoye and Singh, 2006) as a good alternate to model count data with too many zeros. A zero-inflated discrete model has the probability function of the form

$$P(Y_i = y_i) = \begin{cases} \varphi + (1-\varphi) f(0), & y_i = 0 \\ (1-\varphi) f(y_i), & y_i > 0, \end{cases} \tag{9}$$

where $f(y_i)$ is the generalized Poisson distribution in (5) and $0 \le \varphi < 1$. From (7) and (8), the mean and variance of the zero-inflated GPR model are $\mathrm{E}(Y_i) = (1-\varphi)\mu_i$ and $\mathrm{V}(Y_i) = (1-\varphi)\mu_i[(1+\alpha\mu_i)^2 + \varphi\mu_i]$. Famoye and Singh (2006) proposed the zero-inflated GPR in (9) to model domestic violence data with too many zeros.

Zero-truncated generalized Poisson regression model:
The zero-truncated regression models arise in those situations where there is no zero by nature of the data. An example for instance is the length of stay at a hospital. Once you are admitted, it is deemed that you have spent at least one day in the hospital. Thus the zeros cannot be observed in this case for all patients admitted into the hospital. For a random variable *Y* with a discrete distribution, where the value of *Y* = 0 cannot be observed, then the zero-truncated random variable Y_t has the probability mass function

$$P(Y_t = y) = P(Y = y) / P(Y > 0)\,,\ y = 1, 2, 3, \ldots \tag{10}$$

For the zero-truncated Poisson, with parameter μ; $P(Y > 0) = 1 - P(Y = 0) = 1 - e^{-\mu}$.
Hence, the probability mass function of zero-truncated Poisson random variable Y_t is

$$P(Y_t = y) = \frac{\mu^y e^{-\mu}}{y!(1-e^{-\mu})}\,,\ y = 1, 2, 3, \ldots \tag{11}$$

Note that the mean of zero-truncated Poisson model is not equal to μ, it is actually $\mu/(1-e^{-\mu})$. Thus, if μ is the mean of an un-truncated model, the mean of a zero-truncated model is given by $\mu / P(Y > 0)$. The variance of zero-truncated Poisson model is

$$\mathrm{V}(Y) = \mu/(1-e^{-\mu}) - \mu^2 e^{-\mu}/(1-e^{-\mu})^2 < \mathrm{E}(Y)\,.$$

Hence, a truncated Poisson model is always under-dispersed.

The probability mass function for the zero-truncated generalized Poisson distribution can be similarly defined to obtain

$$P(Y_t = y) = \frac{\theta^y (1+\alpha y)^{y-1} e^{-\theta(1+\alpha y)}}{y!(1-e^{-\theta})}\,,\ y = 1, 2, 3, \ldots$$

The mean and variance of zero-truncated generalized Poisson distribution are, respectively,

$$\mathrm{E}(Y_t) = \theta(1-\alpha\theta)^{-1}(1-e^{-\theta})^{-1}$$

and $\mathrm{V}(Y_t) = \theta(1-\alpha\theta)^{-3}(1-e^{-\theta})^{-1} - \theta^2(1-\alpha\theta)^{-2} e^{-\theta}(1-e^{-\theta})^{-2}$.

The truncated generalized Poisson distribution can be used in count data regression when the data is truncated at point zero. One can show that the truncated model satisfy the properties of over-, equi-, and under-dispersion. One can assume that the parameter θ is a function of the predictors x_i.

Censored generalized Poisson regression model:
For some observations in a data set, the value of Y_i may be censored. If no censoring occurs for the i^{th} observation, $Y_i = y_i$. However, if censoring occurs for the i^{th} observation, we know that Y_i is at least equal to y_i i.e. $Y_i \ge y_i$. We now have

$$P(Y_i \ge y_i) = \sum_{j=y_i}^{\infty} P(Y = j) = \sum_{j=y_i}^{\infty} f(\mu_i, j) = 1 - \sum_{j=0}^{y_i - 1} f(\mu_i, j) = P(\mu_i, j), \qquad (12)$$

where $f(\mu_i, y_i) = f(y_i)$ is the ordinary GPR model defined in (5).

Define an indicator variable d_i as

$$d_i = \begin{cases} 1, & Y_i \ge y_i \\ 0, & \text{otherwise.} \end{cases} \qquad (13)$$

The likelihood function of censored generalized Poisson regression (CGPR) model is

$$L(\beta, y_i) = \prod_{i=1}^{n} [f(\mu_i, y_i)]^{1-d_i} [P(\mu_i, y_i)]^{d_i} . \qquad (14)$$

One can take the log-likelihood and use it to find the maximum likelihood estimates.

When α = 0 and the condition $Y_i \ge y_i$ in (13) is replaced with $Y_i \ge C$, where C is a constant, the result in (14) reduces to censored Poisson regression (Terza, 1985) with a constant censoring threshold. Terza (1985) pointed out that this kind of censoring may be imposed on the data by survey design, or it may reflect some theoretical or institutional constraints. If α = 0 and the condition $Y_i \ge y_i$ in (13) is replaced with say $x_i \ge C$ or $x_i \le C$ (where x_i is an explanatory variable) the result in (14) reduces to censored Poisson regression (Caudill and Mixon, 1995) with variable censoring thresholds.

Wang and Famoye (1997) analyzed a data set on fertility from Michigan Panel Study of Income Dynamics (PSID). PSID is a large national longitudinal data set that began in 1968 with approximately 5500 households. The sample has been followed each year since 1968. From the wave in 1989 interviewing year, Wang and Famoye (1997) selected married women aged between 18 and 40 who are not head of households and with nonnegative family income. With this restriction, only 1954 married women were used in the analysis. For the purpose of illustrating censored generalized Poisson regression (CGPR) model in the paper, the restriction on age was dropped and this led to a sample of 2936 married women. The dependent variable, the total number of children up to 17 years old in a family, is a nonnegative integer ranging from zero to nine in the sample.

The mean 1.2922 and variance 1.5016 of the dependent variable are somehow close. This suggests that the data may be equi-dispersed and thus either the Poisson regression model or the GPR model will be adequate for analyzing the data. The purpose of this example is to demonstrate censoring and not to show which independent variable is significant.

Table 1: Parameter estimates for Poisson and censored Poisson regression

	Poisson model	Poisson model for	Censored Poisson
	for complete data	truncated data	for censored data
Parameter	$y_i = 0, 1, \ldots, 9$	$y_i = 0, 1, 2, 3, 4$	$y_i = 0, 1, 2, 3, 4$
β_0	2.0686 ± 0.1511	2.0200 ± 0.1520	2.0711 ± 0.1528
β_1	–0.2657 ± 0.0356	–0.2524 ± 0.0359	–0.2629 ± 0.0360
β_2	–0.0193 ± 0.0041	–0.0181 ± 0.0041	–0.0177 ± 0.0041
β_3	–0.1226 ± 0.0651	–0.0974 ± 0.0655	–0.1055 ± 0.0656
β_4	–0.2811 ± 0.0379	–0.2729 ± 0.0382	–0.2846 ± 0.0382
β_5	0.3057 ± 0.0575	0.2973 ± 0.0576	0.3026 ± 0.0577
β_6	–0.0050 ± 0.0087	–0.0037 ± 0.0088	–0.0051 ± 0.0088
β_7	0.0035 ± 0.0071	0.0016 ± 0.0072	0.0023 ± 0.0072
β_8	–0.0143 ± 0.0187	–0.0135 ± 0.0188	–0.0132 ± 0.0188
β_9	–0.0211 ± 0.0038	–0.0225 ± 0.0039	–0.0230 ± 0.0039
β_{10}	–0.0147 ± 0.0066	–0.0149 ± 0.0066	–0.0154 ± 0.0067
β_{11}	0.0118 ± 0.0078	0.0130 ± 0.0078	0.0129 ± 0.0079
β_{12}	–0.0545 ± 0.0340	–0.0545 ± 0.0342	–0.0593 ± 0.0342
Pearson χ^2	2936.78	2781.22	2752.82
Log-likelihood	–4039.00	–3990.55	–3992.00

Table 2: Parameter estimates for GPR and censored GPR models

	GPR model for	GPR model for	Censored GPR
	complete data	truncated data	for censored data
Parameter	$y_i = 0, 1, \ldots, 9$	$y_i = 0, 1, 2, 3, 4$	$y_i = 0, 1, 2, 3, 4$
β_0	2.0549 ± 0.1488	1.9745 ± 0.1430	2.0542 ± 0.1510
β_1	–0.2665 ± 0.0350	–0.2554 ± 0.0336	–0.2620 ± 0.0353
β_2	–0.0188 ± 0.0041	–0.0166 ± 0.0039	–0.0173 ± 0.0041
β_3	–0.1228 ± 0.0643	–0.0976 ± 0.0624	–0.1048 ± 0.0648
β_4	–0.2797 ± 0.0371	–0.2680 ± 0.0355	–0.2822 ± 0.0375
β_5	0.3047 ± 0.0567	0.2935 ± 0.0552	0.3010 ± 0:0570
β_6	–0.0054 ± 0.0086	–0.0049 ± 0.0083	–0.0054 ± 0.0087
β_7	0.0034 ± 0.0070	0.0015 ± 0.0067	0.0022 ± 0.0071
β_8	–0.0139 ± 0.0183	–0.0121 ± 0.0176	–0.0129 ± 0.0185
β_9	–0.0211 ± 0.0038	–0.0223 ± 0.0037	–0.0229 ± 0.0038
β_{10}	–0.0146 ± 0.0065	–0.0147 ± 0.0064	–0.0152 ± 0.0066
β_{11}	0.0118 ± 0.0077	0.0130 ± 0.0074	0.0129 ± 0.0077
β_{12}	–0.0541 ± 0.0334	–0.0567 ± 0.0321	–0.0587 ± 0.0336
α	–0.0117 ± 0.0096	–0.0404 ± 0.0096	–0.0115 ± 0.0111
Pearson χ^2	2933.37	2769.96	2751.80
Log-likelihood	–4038.29	–3982.93	–3988.39

Famoye and Wang (2004) used both the Poisson regression and generalized Poisson regression (GPR) models to analyze the complete data set without any censoring. About 4.22% of the samples have dependent variable $y_i \geq 4$. To see the effects of censoring on the data, they took the values of all y_i's greater than or equal to 4 to be exactly equal to 4 (Truncated data). The new data was analyzed by using both the standard Poisson regression and standard GPR models. In the analyses, the complete data was assumed to have y_i = 0, 1, 2, 3, 4. Finally, all values of $y_i \geq 4$ were considered as censored and censored Poisson regression and CGPR models were applied to fit the data. The parameter estimates with their standard errors under the Poisson and censored Poisson regression models are presented in Table 1. The corresponding results for the GPR and CGPR models are given in Table 2.

Column 2 of Tables 1 and 2 contains the parameter estimates when the complete data (i.e. y_i = 0, 1, 2, …, 9) is analyzed. The parameter estimates in column 3 of the tables are the results obtained after setting all values of $y_i \geq 4$ to 4 and analyzing the truncated data with ordinary Poisson regression and ordinary GPR models. The estimates in columns 2 and 3 are somehow different especially for the GPR model in Table 2. The results in column 4 represent the estimates from censored Poisson regression and CGPR models. The estimates in column 4 are much closer to the results in column 2. The implication is that analyzing the truncated data without taking into consideration the censoring will lead to inefficient estimates. The asymptotically normal Wald type "t"-values for testing the significance of parameter α in CGPR are respectively −1.22, −4.19, and −1.04 for columns 2, 3, and 4. It is interesting to note that parameter α is significant only in column 3. Thus, the complete data and the truncated data gave conflicting results when standard regression models are used for analysis whereas the complete data and the censored data gave similar results when censored models are used to analyze the censored data. This analysis supports the point that censoring should be taken into consideration when a censored data is used.

Categorized generalized Poisson regression model:
Basu and Famoye (2004) suggested a count data model for the number of violent incidents and it is given by

$$y_i = g(w_i, z_i, \beta) + e_i,\ i = 1, 2, 3, \ldots, n, \qquad (15)$$

where y_i is the number of violent incidents encountered by the i^{th} person over a time period, w_i is the set of economic variables such as income of individual i, z_i is the set of non-economic variables such as age of individual i, β is a vector of unknown parameters and e_i is an unobservable disturbance term. One can assumed that y_i, the number of violent incidents, is a generalized Poisson random variable in (5), and the expected number of violent incidents is given by

$$\mu_i = \mathrm{E}(Y_i \mid x_i) = \exp(X_i'\beta), \qquad (16)$$

where x_i is the vector of explanatory variables (w_i, z_i).

The data analyzed by Basu and Famoye (2004) are grouped (or categorized) for values of $y_i \geq 3$. A substantial amount of information in a data set is lost when some categories (or groups) are combined. To analyze categorized (grouped) dependent variable, one should specify all category probabilities based on the regression model. Suppose the count dependent variable y_i is partitioned into s categories with probabilities P_{ij}, i = 1, 2, …, n, j = 1, 2, …, s, where

$$P_{ij} = \Pr(l_j \leq y_i \leq u_j) = \sum_{r=l_j}^{u_j} f(\mu_i, r),\ \ j = 1,\ 2,\ \ldots,\ s-1,$$

and

$$P_{is} = \Pr(y_i \ge l_s) = 1 - \sum_{r=0}^{u_{s-1}} f(\mu_i, r) .$$

Define an indicator variable d_{ij} as

$$d_{ij} = \begin{cases} 1, \text{ if the } y_i \text{ is in the } j^{\text{th}} \text{ category} \\ 0, \text{ otherwise.} \end{cases}$$

The log-likelihood function for categorized generalized Poisson regression model can be written as

$$\log L(\alpha, \beta;\, y_i) = \sum_{i=1}^{n} \sum_{j=1}^{s} d_{ij} \log P_{ij} . \tag{17}$$

The dependent variable, the total number of violent incidents of husband to wife, is partitioned into the seven categories y_i = 0, y_i = 1, y_i = 2, 3 ≤ y_i ≤ 5, 6 ≤ y_i ≤ 10, 11 ≤ y_i ≤ 20, and y_i ≥ 21. This partition is the same as the coding in the observed data.

The data used in this paper are from the Department of Justice's National Crime Survey (ICPSR 7733), which is explained in Straus and Gelles (1976). In the analysis the dependent variable, violence, represents the number of violent behavior of husband towards wife over one year period. The seven independent variables in Table 3 are conflict (a conflict index), depend (an economic dependence index), agediff (the age difference between the partners), years (total number of years married or living together), income (the total family income before taxes) and education difference, which is husband's education level minus the wife's education level. The education difference is categorized into three levels 'less than zero – wife is more educated', 'zero – couples have the same education', and 'more than zero – wife is less educated'. Two dummy variables are used to represent education difference. The dummy variables are wife_l (the wife is less educated) and wife_m (the wife is more educated). More information on these variables can be found in Basu and Famoye (2004).

Table 3: Effects of Economic Dependence on Incidents of Violence- Comparison among ordinary Poisson, ordinary GPR, and categorized GPR

	Ordinary Poisson	Ordinary GPR	Censored GPR
Variable	Estimate ± se	Estimate ± se	Estimate ± se
Constant	1.3984 ± .1059*	0.8202 ± .5618	0.9879 ± .6864
Conflict	0.8937 ± .0239*	1.1239 ± .1558*	1.2445 ± .1889*
Dependence	0.0595 ± .0161*	0.1539 ± .0990	0.1991 ± .1170
Age difference	–0.0014 ± .0047	–0.0486 ± .0224*	–0.0633 ± .0269
Wife is less educated	0.0875 ± .0399*	–0.1935 ± .2194	–0.0797 ± .2619
Wife is more educated	–0.2090 ± .0497*	–0.8023 ± .2416*	–0.8118 ± .2809*
Years living together	–0.0648 ± .0021*	–0.0758 ± .0090*	–0.0880 ± .0107*
Income	–0.3032 ± .0069*	–0.2676 ± .0372*	–0.2952 ± .0440*
Alpha (α)		2.6272 ± 0.1681*	2.7049 ± 0.1797*
Log-likelihood	–6155.74	–1864.09	–1462.65

* significant at 0.05

In Table 3, the dispersion parameter is positive and significant, indicating that ordinary Poisson regression model is not appropriate. In comparing the log-likelihood values for the ordinary Poisson, ordinary GPR and categorized GPR in Tables 3, we notice that the categorized GPR has the best (largest) value and hence, categorized GPR is the most appropriate model to apply to fit the number of violent incidents. In ordinary generalized Poisson regression model, the values 4, 8, 15, and 25 are respectively used for the classes 3-5, 6-10, 11-20, and 21+. In categorized GPR model, the probabilities of all values in each category are taken into account during estimation. Thus, a categorized model is more accurate and provides a better fit to the data.

5. Comparison tests for non-nested models

In this section, we wish to compare model $f(y_i)$ with model $g(y_i)$. Given two regression models, we consider the hypothesis

$$H_0 : \text{model } f(y_i) \text{ and model } g(y_i) \text{ are equivalent} \tag{18}$$

against

$$H_f : f(y_i) \text{ is better than } g(y_i) \text{ or } H_g : g(y_i) \text{ is better than } f(y_i). \tag{19}$$

The likelihood ratio statistic for testing model $f(y_i)$ against model $g(y_i)$ is defined as

$$L_* = \sum_{i=1}^{n} \log\left(\frac{f(y_i)}{g(y_i)}\right). \tag{20}$$

If the two models $f(y_i)$ and $g(y_i)$ are nested, the statistic in (20) will follow a chi-square distribution. If the two models are not nested, the statistic in (20) is not chi-square distributed.

Vuong (1989) used the Kullback-Liebler Information Criterion to discriminate between two non-nested models. To test the null hypothesis H_0 in (18), Vuong proposed the test statistic

$$Z_* = \frac{L_*}{\hat{\omega}\sqrt{n}}, \tag{21}$$

where $\omega^2 = \frac{1}{n}\sum_{i=1}^{n}\left[\log\left(\frac{f(y_i)}{g(y_i)}\right)\right]^2 - \left[\frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{f(y_i)}{g(y_i)}\right)\right]^2$, is an estimate of the variance of $L_* / \sqrt{n}$. For a non-nested model, Vuong (1989) showed that Z_* is approximately standard normal distributed under the null hypothesis that models $f(y_i)$ and $g(y_i)$ are equivalent. At significant level α, the null hypothesis is rejected in favor of H_g if $Z_* < -z_{\alpha/2}$ and the null hypothesis is rejected in favor of H_f if $Z_* > z_{\alpha/2}$. However, if $|Z_*| \leq z_{\alpha/2}$, we fail to reject the null hypothesis. Thus, we are unable to say that models $f(y_i)$ and $g(y_i)$ are not equivalent.

Clarke (2007) proposed a non-parametric alternative to the Vuong's test. The test by Clarke is distribution-free and it is based on a modified paired sign test on the differences in the individual log-likelihood values from the two non-nested models $f(y_i)$ and $g(y_i)$. The null hypothesis of this non-parametric test is equivalent to

$$H_0 : P\left[\log\left(\frac{f(y_i)}{g(y_i)}\right) > 0\right] = 0.5. \qquad (22)$$

Thus, about half the log-likelihood ratios should be greater than 0 and half should be less than 0. The test statistic is based on the number D^+ of positive differences between the log-likelihood values where D^+ is binomial distributed with parameter *n* and probability 0.5 under the null hypothesis.

While the Clarke's test determines whether the median log-likelihood ratio is statistically different from zero, the Vuong's test determines whether the average log-likelihood ratio is statistically different from zero. Since D^+ is binomial under H_0, we have $E(D^+) = n/2$ and $V(D^+) = n/4$. If model $f(y_i)$ is better than model $g(y_i)$, D^+ will be significantly greater than *n*/2, its expected value under H_0. To test the null hypothesis in (18) against the alternative in (19) at significant level α, we compute the numbers D^+ of positive and D^- of negative differences between the log-likelihood values. Then, we compute the *p*-values for both D^+ and D^-. If the *p*-value for D^+ is below $\alpha/2$, we reject H_0 in favor of model $f(y_i)$ is better than model $g(y_i)$. However, if the *p*-value for D^- is below $\alpha/2$, we reject H_0 in favor of model $g(y_i)$ is better than model $f(y_i)$. When both *p*-values for D^+ and D^- exceed $\alpha/2$, we fail to reject the null hypothesis of equivalence.

According to Clarke (2007), the Vuong's test considers the degree to which the log-likelihood ratio exceeds zero. However, the distribution-free test does not consider the degree to which the log-likelihood ratio exceeds zero, but only if the ratio is positive or negative. Thus, some valuable information may be ignored by the distribution-free test.

Famoye (2011) used the Clarke's test to compare some bivariate regression models and noticed that the Clarke's test may be too sensitive since it only counts the number of positive and negative differences. It is quite possible that the Clarke's test may be highly significant when the log-likelihood values from the two comparison models are about the same. As an illustration, Famoye (2011) selected one of the data sets that were generated for *n* = 100. The data is examined further on the behavior of Vuong's test and Clarke's test. After fitting the data to all the bivariate models, the log-likelihood statistics are –560.69 for bivariate negative binomial regression (BNBR) model, –559.85 for the bivariate generalized Poisson regression (BGPR) model and –558.60 for the bivariate Poisson log-normal regression (BPLR) model. The Vuong statistics show that all models are equivalent.

The Clarke's *p*-values for both numbers of positive and negative differences show that the BGPR and BNBR models are equivalent. The *p*-values for the number of positive and negative differences are respectively 0.9996 and 0.0009 when BGPR model is compared with BPLR model. Thus, we reject the null hypothesis in (18) in favor of the alternative that the BPLR model is better than the BGPR model. The Clarke's test seems to be too sensitive in showing that BPLR model is better than the BGPR model even though the data is generated from the BGPR model. Famoye (2011) also showed that both the Vuong's test and the Clarke's test performed poorly in achieving the nominal significance level, but the Vuong's test is more powerful than the Clarke's test.

Clarke (2007) showed that the distribution-free test is asymptotically more efficient than the Vuong's test when the distribution of the individual log-likelihood ratios is leptokurtic. Clarke (2007) illustrated this with double exponential distribution which is symmetric. Famoye (2011) noticed that the empirical distribution of the log-likelihood ratios is leptokurtic, but skewed to the right hand side especially for large sample sizes. Furthermore, the Vuong's test appears to be more powerful than the distribution-free test when comparing the BGPR and BNBR models. Thus, the Clarke's test may not be appropriate for comparing the bivariate regression models. Further work is needed to develop a more powerful test to compare these bivariate regression models.

Famoye (2005) conducted a simulation study to compare the generalized Poisson and negative binomial regression models when the true data generating process exhibits over-dispersion. The Vuong's test was used in the study. In general, the GPR model can be used in place of the NBR model as both models are equivalent with a high percentage. In small data sets (n = 25 to n = 200), the GPR model has an advantage over the NBR model when the data has a high proportion of zeros. In addition to the fact that the GPR can model under-dispersion, it appears to be a model that one should always apply. In terms of estimation, one model is not easier to estimate than the other. In fact, in the simulation study, there are a few cases when the NBR model failed to converge in data generated from the GPR model. This is not observed for the GPR model when the data is generated from the NBR model.

6. Conclusion

The generalized Poisson regression model and its modification are discussed in the paper. Some of the modified Poisson regression models may be applicable to one type of dispersed data while the modified generalized Poisson regression model is, in general, applicable to over-, equi- or under-dispersed data. Quite often, we do not have equi-dispersed data. For example, it is hard to know the type of dispersion exhibited by a censored data. Hence, a censored generalized Poisson regression model that can accommodate any kind of dispersion should be applied. The modified GPR model is more versatile than the modified Poisson regression or even modified negative binomial regression models. Based on the simulation studies and comparison tests, we recommend the use of ordinary or modified generalized Poisson regression model for any count data.

References

Basu, B. and Famoye, F. (2004) Domestic violence against women, and their economic dependence: A count data analysis, *Review of Political Economy*, 16(4), 457-472.

Clarke, K.A. (2007) A simple distribution-free test for nonnested model selection. *Political Analysis*, 15, 347-363.

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

Caudill, S.B., Mixon Jr., F.G. (1995) Modeling household fertility decisions: estimation and testing censored regression models for count data, *Empirical Economics*, 20, 183–196.

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

Famoye, F. (2011) Comparisons of some bivariate regression models. *Journal of Statistical Computation and Simulation*. [DOI:10.1080/00949655.2010.543679]

Famoye, F. (2005) Count data modeling: Choice between generalized Poisson model and negative binomial model, *Journal of Applied Statistical Science*, 14(1-2), 99-106.

Famoye, F. (1993) Restricted generalized Poisson regression model, *Communications in Statistics-Theory and Methods*, 22(5), 1335-1354.

Famoye, F. and Singh, K.P. (2006) Zero inflated generalized Poisson regression model with applications to domestic violence data, *Journal of Data Science*, 4(1), 117-130.

Famoye, F. and Singh, K.P. (2003) On inflated generalized Poisson regression models, *Advances and Applications in Statistics*, 3(2), 145-158.

Famoye, F. and Wang, W. (2004) Censored generalized Poisson regression model, *Computational Statistics & Data Analysis*, 46, 547-560.

Hilbe, J.M. (2007) *Negative Binomial Regression*, Cambridge University Press, Cambridge, UK.

Lawless, J.F. (1987) Negative binomial and mixed Poisson regression, *Canadian J. of Statistics*, 15(3), 209-225.

Straus, M.A. and Gelles, R.J. (1976) *Physical Violence in American Families* ([computer file], Conducted by Straus MA, University of New Hampshire, and Gelles, R.J., University of Rhode Island, 1976. 2nd ICPSR ed. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [producer and distributor]).

Terza, J.V. (1985) A tobit-type estimator for the censored Poisson regression model, *Economic Letters*, 18, 361–365.

Vuong, Q.H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, 57(2), 307-333.

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