A New Generalized Poisson Regression Model for Count Data with Multiple Dispersions

J. O. Oyekunle[1]; W. B. Yahya[2]

[1]Department of Statistics,
School of Applied Sciences,
Federal Polytechnic, Ede, Nigeria.

[2]Department of Statistics,
University of Ilorin,
Ilorin, Nigeria
E-mail: oyekunle.olufunmike@yahoo.com[1]

***Abstract* — In this paper, a modified Generalized Poisson Regression (mGPR) for modelling count data with more than one heterogeneous dispersed data condition is presented. The mGPR is largely an extension of the GPR model (with one dispersion parameter) that is capable of detecting the presence of either the underdispersion or overdispersion in count data but not both if such are present or dominant in the data. The new mGPR contains two dispersion parameters that are capable of detecting if the data are plagued with the two dispersed data structures but with one of them more dominant in the data than the other as often the case with many real-life data situations. The parameters of the mGPR were estimated using the maximum likelihood method. Results from Monte-Carlo studies showed that the mGPR model is more efficient than the classical GPR model on count data with more than one dispersed structure as evident from the results of the Deviance and log-likelihood. Like the GPR model, the mGPR model reduces to the classical Poisson regression model if the data contain equidispersed structure. Results from the Monte-Carlo studies were validated on real-life data.**

Keywords - *Generalized Poisson Regression, Modified Generalized Poisson Regression, equi-dispersion, under-dispersion, over-dispersion, count data.*

I. Introduction

The benchmark model for count data is the Poisson distribution. Count data is a type of dataset in which the observations can take only the non-negative integer values (0, 1, 2, 3, …) and these integers arise from counting rather than ranking. Statistical analysis involving count data can take several forms depending on the context in which the data arise. Simple counts can be taken as the number of occurrences of a particular event in a month for several years. Categorical data on the other hand represents the number of items belonging to each of several categories.

The Poisson distribution is a particular case of the generalized linear model, in which the conditional distribution of the dependent variable follows a Poisson law and the link function is logarithmic (Winkelmann et. al. 1994, Trussel and Rodriguez, 1990, Cameron et. al. 1998). It presents several advantages for statistical analysis of fertility as noted by Schoumaker (2006). Poisson regression estimates the effects of explanatory variables on rates; the logarithmic form of the model is such that the exponents of the regression coefficients represent the relationships between the rates as it applies to different groups of individuals. For example, the fertility rates of different groups of women.

One restriction of the Poisson distribution is that it allows only a single parameter to estimate the mean and variance in which the variance is equal to the mean (equidispersion) which is often than not is not sustainable in real-life situations. The cases of overdispersion (variance > mean), under-dispersion (variance < mean), and excess zeros in which the observed data contained more zero counts than what is ordinarily expected of Poisson distributed data are some of the scenarios that exist in many practical situations. The imposition of the Poisson regression model on such data may lead to biased parameter estimates, false conclusions, and wrong decisions. Famoye, Wulu, and Singh (2001) noted that the Poisson Regression model is not appropriate when a dataset exhibits over-dispersion.

Separate models have been proposed in the literature to handle each of these cases.

The situation in which a count dataset contains two of the dispersed data structures, especially if the data exhibits the presence of both overdispersion and underdispersion is yet to be addressed in the literature. This kind of data structure abounds in many real-life situations. An example of this can be found in the fertility dataset (number of children born) that was stratified by either the levels of education or religion. Sub-populations with high and low levels of educations will obviously have different fertility structures which may trigger different dispersion structures in the data. Imposing either the Poisson or any of the Poisson-related models on such data might yield a suboptimal model. This kind of data situation can best be captured by a model that considered a possible presence of more than one dispersion structure in the data as is the case with the proposed mGPR model in this work.

II. The Poisson Regression Model

The general form of the Poisson distribution is given as

$$P(y_i) = \frac{\lambda^y \exp(-\lambda)}{y_i!} \ ; \quad y = 0,1,2,\ldots \qquad (1)$$

where λ is the parameter of Poisson distribution and it is a function of some explanatory variables, X, which takes the exponential form below

$$\lambda_i = \exp(X_i\beta)$$

The dependent variable y_i represents rate, for example, the number of births occurring to a woman (i) over a given period. The conditional mean of y_i that is; $E(y_i) = \exp(X_i\beta)$which also denotes the conditional variance of y_i since the Poisson distribution is equi-dispersed.

III. The Generalized Poisson Regression Model

Consul (1989) first presented the generalized Poisson regression in a monograph. Further references are; Consul and Famoye(1992), Famoye(1993) and Wang and Famoye(1997). The latter references introduced exogenous variables and thus a generalized Poisson regression model (GPR). Santos and Silva(1997b) extended the model to truncated data. The generalized Poisson is a genuine alternative to the generalized event count model as it allows for both over-and underdispersion and nests the classical Poisson regression model as a special case. This is achieved by introducing one additional parameter θ. The probability distribution function was then written as (Consul, 1989)

$$f(y) = \begin{cases} \dfrac{\theta(\theta + y\gamma)^{y-1} e^{-\theta - y\gamma}}{y!}, & y = 0,1,2,\ldots \\ 0 \ for \ y > m\,, & when \ \gamma < 0 \end{cases} \qquad (2)$$

Where$\theta > 0, max\left[-1, -\frac{\theta}{m}\right] < \gamma \leq 1$ and$m(\geq 4)$ is the largest positive integer for which
$\theta + m\gamma > 0$ when γ is negative.
The generalized Poisson model uses the following reparameterizations:

$$\theta_i = \frac{\lambda_i}{1 + \alpha\lambda_i}$$

$$\gamma_i = \frac{\alpha\lambda_i}{1 + \alpha\lambda_i}$$

where $\lambda_i = \exp(x_i'\beta)$.
Let the response variable Y$_i$, representing count (number of children born to a woman) be a generalized Poisson random variable, Famoye (1993). It was later referenced by Wang and Famoye (1997) and was used to model household fertility decisions. The probability function of $Y_i$is given by

$$f_i(y_i; \mu_i, \alpha) = \left(\frac{\mu_i}{1 + \alpha\mu_i}\right)^{y_i} \frac{(1 + \alpha y_i)^{y_i - 1}}{y_i!} exp\left(\frac{-\mu_i(1 + \alpha y_i)}{1 + \alpha\mu_i}\right), \quad y_i = 0,1,2,.. \qquad (3)$$

$$\mu_i = \mu_i(x_i) = \exp(x_i\beta) = \exp(\beta_0 + \beta_1 x_{1i})$$

Where x$_i$ is a (k-1) dimensional vector of explanatory variables such as educational level and some other personal characteristics of couples in a family as well as some demographic attributes of the family, and β is a *k*-dimensional vector of regression parameters.

The mean and variance of Y_i is given by

$$E(Y_i|x_i) = \mu_i \qquad (4)$$

and

$$V(Y_i|x_i) = \mu_i(1 + \alpha\mu_i)^{\,2} \qquad (5)$$

The generalized Poisson model in (3) is a natural extension of the standard Poisson model. When α, which is the dispersion parameter equals to zero then equation (1) reduces to the Poisson probability function and$E(Y_i|x_i) = V(Y_i|x_i)$(equi-dispersion). For $\alpha > 0$, $V(Y_i|x_i) > E(Y_i|x_i)$and the generalized Poisson regression model in (1) represents count data with over-dispersion. For $\alpha > 0$, $V(Y_i|x_i) < E(Y_i|x_i)$ and the model in (1) represents count data with under-dispersion. The dispersion parameter can be estimated simultaneously with the coefficients in the GPR model in (1) as stated in Famoye(1997)

The likelihood function of the GPR model (1) is given by

$$L(\alpha,\beta;y_i) = \sum_{i=1}^{n}\left\{\left(\frac{\mu_i}{1+\alpha\mu_i}\right)^{y_i}\frac{(1+\alpha y_i)^{y_i-1}}{y_i!}exp\left(\frac{-\mu_i(1+\alpha y_i)}{1+\alpha\mu_i}\right)\right\}$$

To estimate (β,α) in the GPR model (3), we first write the log-likelihood function, $LnL(\alpha,\beta;y_i)$as follows:

$$LnL(\alpha,\beta;y_i) = \sum_{i=1}^{n} log\left\{\left(\frac{\mu_i}{1+\alpha\mu_i}\right)^{y_i}\frac{(1+\alpha y_i)^{y_i-1}}{y_i!}exp\left(\frac{-\mu_i(1+\alpha y_i)}{1+\alpha\mu_i}\right)\right\}$$

$$= \sum_{i=1}^{n} y_i log\left(\frac{\mu_i}{1+\alpha\mu_i}\right) + (y_i-1)log(1+\alpha y_i) - \frac{\mu_i(1+\alpha y_i)}{1+\alpha\mu_i} - \log(y_i!) \qquad (8)$$

The maximum likelihood equations for estimating $\alpha$and β are obtained by taking the partial derivatives of the above equation (4) and equating it to zero. Thus, we have;

$$\frac{\partial LnL}{\partial\alpha} = \sum_{i=1}^{n}\left\{y_i\frac{\partial}{\partial\alpha}log\left(\frac{\mu_i}{1+\alpha\mu_i}\right) + (y_i-1)\frac{\partial}{\partial\alpha}log(1+\alpha y_i) - \frac{\partial}{\partial\alpha}\left[\frac{\mu_i(1+\alpha y_i)}{1+\alpha\mu_i}\right] - \frac{\partial}{\partial\alpha}\log(y_i!)\right\} = 0 \qquad (9)$$

Since $\frac{\partial}{\partial\alpha}\log(y_i!) = 0$

$$\frac{\partial LnL}{\partial\alpha} = \sum_{i=1}^{n}\left\{\frac{-y_i\mu_i}{1+\alpha\mu_i} + \frac{y_i(y_i-1)}{1+\alpha\mu_i} - \frac{\mu_i(y_i-\mu_i)}{(1+\alpha\mu_i)^2}\right\} = 0 \qquad (10)$$

$$\frac{\partial LnL}{\partial\beta_r} = (6) \frac{\partial LnL}{\partial\mu_i} * \frac{\partial\mu_i}{\partial\beta_r} = 0$$

IV. The Proposed Modified GPR (mGPR) Model

Due to (7) the fact that certain count data, such as fertility data are heterogeneous in nature, care must be taken in selecting a model that can be used to model it in order to reflect the heterogeneity in such data. This explains the motivation behind this research in which the focus is to propose a single model that would efficiently fit the count dataset with an inherent heterogeneous structure like equi-dispersed, over-dispersed and under-dispersed structures simultaneously. Our proposed model was formulated by extending the GPR in (3) to include two dispersion parameters α_1 and α_2 in place of α to account for the possible existence of such heterogeneous data structure. This new proposal is called a Modified Generalized Poisson Regression (mGPR) model. (Oyekunle and Yahya, 2016)

Assuming we have three different data structures (as may be induced by say, educational levels) in a sample of size (n) with group sample sizes $m_1,$ m_2 and $m_3,$ m_1+m_2+$m_3 = n$. For instance, the educational levels of respondents can have distinct structures of high, low, and no levels of education and it is expected that the fertility preference (number of children desired) could be influenced by these three structures of respondents' educational levels. Such ideal data structure is illustrated in Table 1.

Table 1: Table of the data structure showing the educational level of respondents and the expected dispersion patterns.

S/N	Level of Education in the sub-population	Fertility	Mean no. of Children in the sub-population	Sub-population Variance	Remark
1 ⋮ m₁	High	Low	μ_1	σ_1	$\mu_1 > \sigma_1$ Under-dispersion
m₁+1 ⋮ m₂	Low	Moderate	μ_2	σ_2	$\mu_2 = \sigma_2$ Equi-dispersion
m₂+1 ⋮ n	No education	High	μ_3	σ_3	$\mu_3 < \sigma_3$ Over-dispersion

Source: Researcher's Conceptualisation, 2013.

Therefore, following the GPR model in (3), the proposed mGPR is given by

$$f_i(y_i; \mu_i, \alpha_1, \alpha_2) = \left(\frac{\mu_i}{1 + (\alpha_1 - \alpha_2)\mu_i}\right)^{y_i} \frac{[1 + (\alpha_1 - \alpha_2)y_i]^{y_i - 1}}{y_i!} exp\left(\frac{-\mu_i(1 + (\alpha_1 - \alpha_2)y_i)}{1 + (\alpha_1 - \alpha_2)\mu_i}\right), \quad y_i; 0,1,2. \qquad (12)$$

where Y_i, μ_i, x_i and β remained as defined earlier and$\mu_i = \mu_i(x_i) = \exp(\beta_0 + \beta_i x_i)$

The mean and variance of Y_i are respectively given by

$E(Y_i|x_i) = \mu_i = \exp(\beta_0 + \beta_i x_i)$ and (13)

$$V(Y_i|x_i) = \mu_i = \mu_i[1 + (\alpha_1 - \alpha_2)\mu_i]^2 = \exp(\beta_0 + \beta_i x_i)\,[1 + (\alpha_1 - \alpha_2)\exp(\beta_0 + \beta_i x_i)]^2 \qquad (14)$$

For null model, $x_i = 0$, hence, (13) and (14) becomes

$$E(Y_i|x_i = 0) = \mu_i = \exp(\beta_0) \qquad (15)$$

$$V(Y_i|x_i = 0) = \mu_i = \mu_i[1 + (\alpha_1 - \alpha_2)\mu_i]^2 = \exp(\beta_0)\,[1 + (\alpha_1 - \alpha_2)\exp(\beta_0)]^2 \qquad (16)$$

respectively.

The modified generalized Poisson model in (12) is an extension of the standard Poisson regression model. When $\alpha_1 = \alpha_2$ the probability function in (12) reduces to the Poisson probability function and $E(Y_i|x_i) = V(Y_i|x_i)$(equidispersion). If $\alpha_1 > \alpha_2$and $E(Y_i|x_i) < V(Y_i|x_i)$; the modified GPR model in (12) represents data with over-dispersion and represents data with under-dispersion if $\alpha_1 < \alpha_2$and $E(Y_i|x_i) > V(Y_i|x_i)$)

The likelihood function of the GPR model (9) is given by:

$$L(\alpha_1, \alpha_2, \beta; y_i) = \sum_{i=1}^{n} \left\{ \left(\frac{\mu_i}{1 + (\alpha_1 - \alpha_2)\mu_i}\right)^{y_i} \frac{[1 + (\alpha_1 - \alpha_2)y_i]^{y_i - 1}}{y_i!} exp\left(\frac{-\mu_i[1 + (\alpha_1 - \alpha_2)y_i]}{1 + (\alpha_1 - \alpha_2)\mu_i}\right) \right\}, \qquad (17)$$

To estimate (β, α1 and α2) we first write the log-likelihood function,$LnL(\alpha_1, \alpha_2, \beta; y_i)$ of the GPR model (12) as:

$$LnL(\alpha_1, \alpha_2, \beta; y_i) = \sum_{i=1}^{n} log \left\{ \left(\frac{\mu_i}{1 + (\alpha_1 - \alpha_2)\mu_i}\right)^{y_i} \frac{[1 + (\alpha_1 - \alpha_2)y_i]^{y_i - 1}}{y_i!} exp\left(\frac{-\mu_i[1 + (\alpha_1 - \alpha_2)y_i]}{1 + (\alpha_1 - \alpha_2)\mu_i}\right) \right\} \qquad (18)$$

$$= \sum_{i=1}^{n} \left\{ y_i\, log\left(\frac{\mu_i}{1 + (\alpha_1 - \alpha_2)\mu_i}\right) + (y_i - 1)\log[1 + (\alpha_1 - \alpha_2)y_i] - \log(y_i!) - \left(\frac{\mu_i[1 + (\alpha_1 - \alpha_2)y_i]}{1 + (\alpha_1 - \alpha_2)\mu_i}\right) \right\} \qquad (19)$$

where,

$$\mu_i = \exp(\beta_0 + \beta_r x_i)\ ; \quad r = 1,2,\ldots,k\,, \quad i = 1,2,\ldots,n$$

The maximum likelihood equation for estimating (β, α_1 and α_2) is obtained by taking partial derivatives of equation (17) with respect to each of the parameters and equating it to zero.

1V. Simulation Study

To compare the proposed mGPR with the existing GPR (Famoye, 1997) model, data were simulated based on the following schemes. Parameter μ, the average live births in the population was set to be 3. In order to impose two heterogeneous groups (e.g. to induce over- and/or under-dispersion) in the data, two groups of sample sizes n1 and n2 were simulated with $n_1 + n_2 = n$. Equi-dispersed data sets were simulated from the classical Poisson distribution while the overdispersed and underdispersed data sets were simulated using the negative binomial distribution. These data sets were combined and mixed in the ratios; 1:3, 3:1, 1:1, 3:2 and 2:3 in an attempt to demonstrate the modeling ability of the models. The dispersion parameter θ was set such that:

$\theta > 0$ for overdispersed and
$\theta < 0$ for underdispersed datasets.

Situations with $n_1 > n_2$, $n_1 < n_2$ and $n_1 = n_2$ were considered. The study generated y-data by considering the following actual sample sizes: $n = 200, 300, 500\ and\ 1000$. Under each situation, we examine the difference between the estimated $(\hat{\alpha}_1 - \hat{\alpha}_2)(for\ mGPR)$ and $\hat{\alpha}$ (for GPR), we also consider their log-likelihood functions and the deviances. Finally, throughout the simulation processes, null model was conjectured.

Two real life data sets: one set of underdispersed data containing take-over bids culled from *countreg* (R. package) and the other set which is overdispersed contains data on recreation demand (number of trips) from *AER* (R. package) were also used to demonstrate the performance of the model in comparism with the existing GPR model. The two data sets were analyzed differently and later combined together to demonstrate the strength of each of the models in identifying the different dispersion structures in the data sets.

Lastly data on Nigerian National Demographic and Health Survey for 1999 consisting of 3552 respondents was used to demonstrate the performance of the model while also comparing this with the existing model. The fertility structure based on the total number of children ever born to a woman after which the data has been partitioned using the women's educational level was used.

All simulations and computations were performed using R statistical package (www.cran.org). The log-likelihoods and deviances for each model were also computed. The estimate of the deviance for the model was taken as:

$$D = -2\left[l(\hat{\mu}_i, y_i) - l(y_i, y_i)\right] \qquad (26)$$

Specifically, the deviances for the mGPR is determined by:

$$D = 2\sum_{i=1}^{n}\left\{ y_i \ln\left[\frac{y_i\left(1+(\hat{\alpha}_1-\hat{\alpha}_2)\mu_i\right)}{\hat{\mu}_i\left(1+(\hat{\alpha}_1-\hat{\alpha}_2)y_i\right)}\right] + \frac{\hat{\mu}_i - y_i}{1+(\hat{\alpha}_1-\hat{\alpha}_2)\mu_i}\right\} \qquad (27)$$

The selection of the best model depends on which models present lower values of Log-likelihood and Deviance.

V. Results and Discussion

Table 2: Dispersion Parameters, Log-likelihood and Deviance of Existing and Proposed Models via Monte-Carlo Simulation (for data mixture 1:3/3:1 and 1:1) n=200

Parameter (n=200)	**NB**			**GPR**			**Modified-GP**		
	50,150	**150,50**	**100,100**	**50,150**	**150,50**	**100,100**	**50,150**	**150,50**	**100,100**
	A_1	**B_1**	**C_1**	**A_2**	**B_2**	**C_2**	**A_3**	**B_3**	**C_3**
$\hat{\alpha}$	18.3	7.65	6.67	**-0.1447**	**0.2094**	**0.0335**	-	-	-
$\widehat{\alpha_1}$	-	-	-	-	-	-	3.4136	2.8063	9.6622
$\widehat{\alpha_2}$	-	-	-	-	-	-	3.5553	2.7686	9.6292
$\widehat{\alpha_1} - \widehat{\alpha_2}$							**-0.1417**	**0.0377**	**0.033**
$\hat{\mu}$	-	-	-	2.4139	46.7044	3.4171	2.4181	3.3394	3.4152
LL	-415.701	-425.235	-427.5535	-13.9033	**-100.557**	-735.5722	-13.9027	**-696.4265**	-735.5715
Deviance	218.2573	226.526	230.9724	961.9202	**2493.752**	145.0967	1337.172	**190.3029**	164.5823

Table 3: Dispersion Parameters, Log-likelihood and Deviance of Existing and Proposed Models via Monte-Carlo Simulation (for data mixture 2:3/3:2) n=200

Parameter (n=200)	**NB**		**GPR**		**Modified-GP**	
	120,80	**80,120**	**120,80**	**80,120**	**120,80**	**80,120**
	D_1	E_1	D_2	E_2	D_3	E_3
$\hat{\alpha}$	9.87	8.88	**0.0385**	**0.0459**	-	-
$\widehat{\alpha_1}$	-	-	-	-	2.8674	1.9501
$\widehat{\alpha_2}$	-	-	-	-	2.8279	1.9025
$\widehat{\alpha_1} - \widehat{\alpha_2}$					**0.0395**	**0.0476**
$\hat{\mu}$	-	-	3.4104	3.4020	3.4133	3.4063
LL	-434.324	-434.787	-739.1573	-738.9322	-739.1563	-738.9317
Deviance	226.1229	225.0118	**145.1708**	**137.9611**	**143.9433**	**137.1727**

Note: A,B,C,D, E represents the sample mix ratios 1:3, 3:1, 1:1, 3:2, 2:3 respectively.

Comparison of the GPR and mGPR

Figure 1

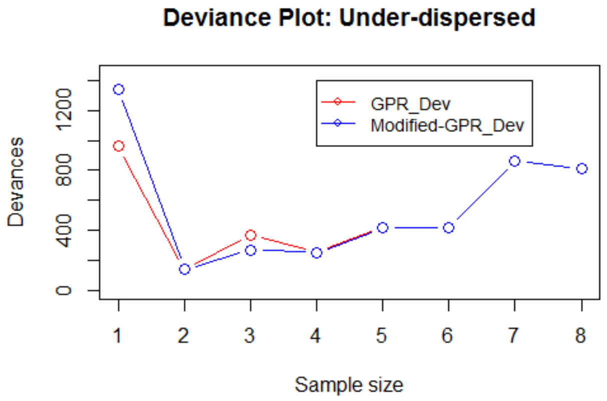

Deviance Plot: Under-dispersed

Figure 2

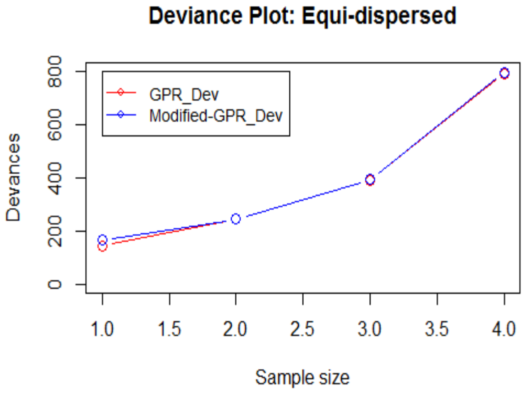
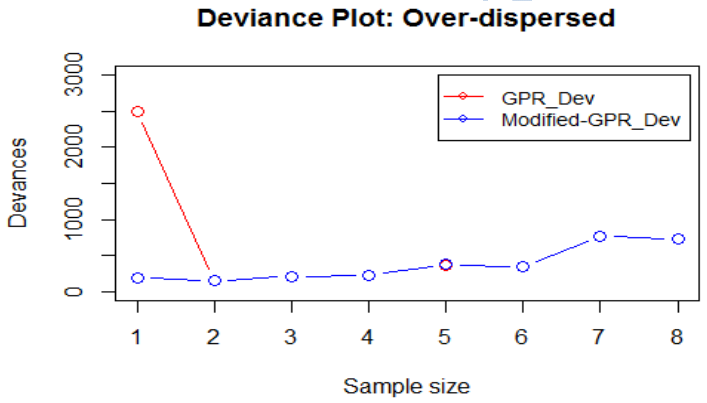

Deviance Plot: Equi-dispersed

Figure 3


Deviance Plot: Over-dispersed

The deviance values for underdispersed, overdispersed and equidispersed simulated datasets across the sample sizes (200,300,500 and 1000) were separately plotted against the sample mix ratios to provide a visual display of the performance of the existing and proposed models as shown in figures 1,2 and 3 above.

The dispersion parameters, Log-likelihood and Deviance for the different data mix were estimated using the NB, GPR and mGPR respectively. Tables 2 and 3 show the estimate for sample size 200 mixed using the mixture ratios described above (1:3, 3:1, 1:1, 3:2, 2:3) while tables 4 and 5 is for sample size 300, 6 and 7 for sample size 500 and 8 and 9 for sample size 1000. It can be seen from the values of the dispersion parameters that the proposed model was able to recognize a set of data as being overdispersed or underdispersed as does the existing model.

The difference here is that the modified (mGPR) model was able to filter the data by separating the data based on the dispersion information present in it by the parameters α_1 and α_2 thereby reflecting the different dispersion structures in the data set. It is also worthy of note that data structure and the quantum of the dispersion information in the data play a prominent role in the efficiency of this series of the count data model.

Comparing columns B_2 and B_3 for GPR and mGPR in Table 2 for sample size 200 mixed in the ratio 1:3 (150,50), it can be seen that the mGPR was able to better estimate the model by having lower log-likelihood and deviance than the existing GPR **(LL = -100.557 and -696.4265; Deviance= 2493.752 and 190.3029)**.

Real-Life Data:

Table 4: Dispersion Parameters, Log-likelihood and Deviance of Existing and Proposed models for Real Life Data (for Under, Over and Combined data sets)

Par	NB			GPR			mGPR		
	Under	Over	Combined	Under	Over	Combined	Under	Over	Combined
	A_1	B_1	C_1	A_2	B_2	C_2	A_3	B_3	C_3
$\hat{\alpha}$	-	-	-	**0.0288**	**1.6712**	**1.1986**	-	-	-
$\widehat{\alpha_1}$	-	-	-	-	-	-	3.8546	0.0174	5.9436
$\widehat{\alpha_2}$	-	-	-	-	-	-	3.8256	-1.6558	4.7466
$\widehat{\alpha_1} - \widehat{\alpha_2}$							**0.029**	**1.6732**	**1.1970**
$\hat{\mu}$	18.6	0.1715	0.2618	1.7407	2.2428	2.1556	1.7328	2.2485	2.1589
LL	-201.1205	-1064.8	-201.1205	-6.1957	**-3961.0**	**-3888.345**	-6.1951	**-3969.0**	**-3888.345**
DD	109.8929	483.78	671.0054	**78.968**	45.134	**83.64808**	**78.795**	51.759	**88.15664**

NB:
Under-dispersed Data
TakeoverBids (variable = bids) from countreg (R. package)

Firms that were targets of takeover bids during the period 1978–1985.A data frame containing 126 observations on 9 variables, out of which **"bids"** which denotes number of takeover bids (after the initial bid received by the target firm).The data were originally used by Jaggia and Thosar (1993), where further details on the variables may be found.
Sources:
Journal of Applied Econometrics Data Archive for Cameron and Johansson (1997).
http://qed.econ.queensu.ca/jae/1997-v12.3/cameron-johansson/

Over-dispersed Data
Recreation Demand (variable = trips) from AER (R. package)

Cross-section data on the number of recreational boating trips to Lake Somerville, Texas, in 1980, based on a survey administered to 2,000 registered leisure boat owners in 23 counties in eastern Texas. A data frame containing 659 observations on 8 variables, out of which **"trips"** denotes the number of recreational boating trips. According to the source (Seller, Stoll and Chavas, 1985, p. 168), the quality rating is on a scale from 1 to 5 and gives 0 for those who had not visited the lake. This explains the remarkably low mean for this variable, but also suggests that its treatment in various more recent publications is far from ideal. For consistency with other sources, we handle the variable as a numerical variable, including the zeros.
Sources:
Journal of Business & Economic Statistics Data Archive.
http://www.amstat.org/publications/jbes/upload/index.cfm?fuseaction=ViewArticles&pub=JBES&issue=96-4-OCT

Combined dataset
Combination of the two datasets (i.e. The under and over-dispersed datasets)

In Table 4, the performance of the models was demonstrated using a set of real data sets as presented above. The two data sets were analyzed as earlier discussed. The negative binomial and the GPR models captured all the data sets as overdispersed: which is expected because they are models commonly used for overdispersed count datasets. The bold figures show the situations when the proposed model demonstrates more efficiency in capturing the dispersion structure in the data sets than the existing models. It could be observed that when the two datasets were combined, the proposed model was found to outperform the existing models in many situations.

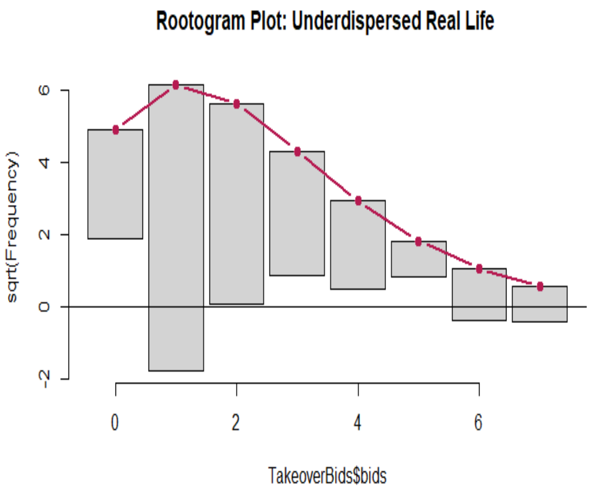


Figure 4

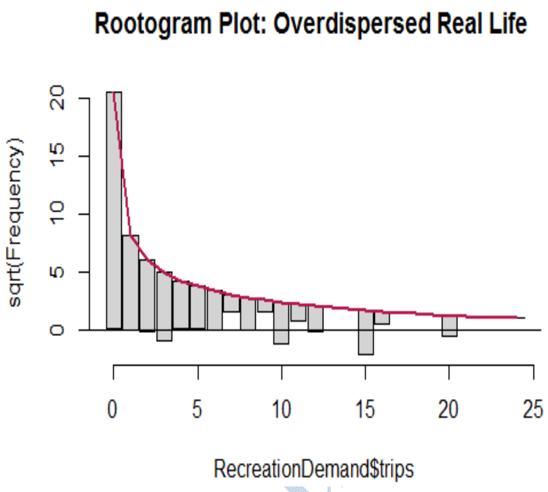


Figure 5

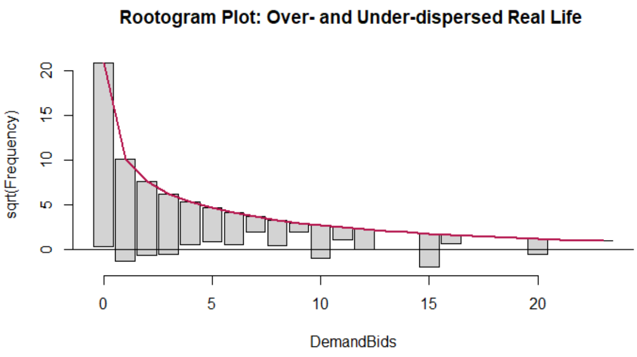


Figure 6

Figures 4,5,6 represents the Rootogram for the Real life data sets viz: underdispersed, overdispersed, and combined sets to summarize the distributional information of the variable used in the data. the vertical axis represents the square root of the frequencies while the horizontal axis represents the response variable as defined.

VI. Conclusion

In this paper, we introduce a modified generalized regression model by extending the existing GPR model with one dispersion parameter α (Famoye, 1993) to have two dispersion parameters α_1 and α_2 , where $\alpha_1 = \alpha_1 - \alpha_2$ in an attempt to really filter make some distinction between the dispersion information present in the data. This modified model is called mGPR and was compared with the existing GPR in order to determine which model will have the strength to recognize the dispersion structure in a given data set with heterogeneous features. The performance of the models were demonstrated using different data sets ranging from simulated data sets; mixed in different ratios to reflect heterogeneity in dispersion patterns, to real life data sets, with heterogeneous dispersion structures. The dispersion parameter(s), log-likelihood and deviance were estimated using the two models. Smaller values of log-likelihood and deviance show the efficiency of any of the models. In many instances most especially when the models were used for real life data sets the mGPR gives log-likelihood and deviance values smaller than the existing GPR model. This shows that the mGPR is more efficient and can be used as an attractive alternative to either the Poisson, the NB and GPR to model overdispersed, equidispersed and overdispersed count data sets simultaneously.

References

1. Cameron A. C, Johansson P (1997). "Count Data Regression Using Series Expansion: With Applications", *Journal of Applied Econometrics*, **12**(3), 203–224.

2. Cameron, A. C., Trivedi, P. K. (1998). *Regression Anaylsis of Count Data.* New York: Cambridge University Press.

3. Cameron AC, Trivedi PK (2013). *Regression Analysis of Count Data*, 2nd ed. Cambridge: Cambridge University Press.

4. Consul, P.C.(1989). *Generalized Poisson Distribution: Properties and Application.* New York; Marcel Dekker.

5. Consul, P.C., Famoye, F. (1992). Generalized Poisson regression model. *Communications in Statistics (Theory and Method). 2(1): 89-109*

6. Famoye F. (1993). Restricted generalized Poisson regression model. *Communication in Statistics – Theory and Methods*. 22:1335-1354.

7. Jaggia S. and Thosar S. (1993). "Multiple Bids as a Consequence of Target Management Resistance: A Count Data Approach", *Review of Quantitative Finance and Accounting*, **3**, 447–457.

8. Oyekunle, J.O. and Yahya W. B. (2016), On Parameters Estimation of Modified Generalised Poisson Regression For Modeling Count Data With Heterogeneous Structures. *Paper presented at the 35th Annual Conference of the Nigerian Mathematical Society (NMS) held at Federal University of Technology, Minna.*

9. Seller, C., Stoll, J.R. and Chavas, J.-P. (1985). Validation of Empirical Measures of Welfare Change: A Comparison of Nonmarket Techniques. *Land Economics*, **61**, 156–175.

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

11. Winkelmann R., Zimmermann K.F. (1994). Count models for demographic data. *Mathematical Population Studies* 4; 205-221.