Models for Count Data in the Presence of Outliers and/or Excess Zero

Usman, M.[1*] and Oyejola, B. A.[2]

1. Federal Polytechnic Bali, Dept of Statistics, P.M.B. 05 Bali, Nigeria
2. University of Ilorin, Dept of Statistics, P.M.B. 1515 Ilorin, Nigeria

*Email: muhdbnuthman2011@gmail.com boyejola2003@yahoo.com

Abstract

Violations of Poisson assumptions usually result in overdispersion, where the variance of the model exceeds the value of the mean. Excess or (deficiency) of zero counts result in overdispersion. Violations of equidispersion indicate correlation in the data, which affect standard errors of the parameter estimates. Model fit is also affected. (Hilbe 2008). Therefore, this study examined the impact of outliers and excess zero on count data in causing overdispersion. The study focus on identifying model(s) which can handle the impact of outliers and excess zero in count data. Datasets based on Poisson model were simulated for sample sizes 20, 50 and 100 and incorporated with outliers and excess zero. Maximum likelihood estimation method was employed in estimating the parameters. Model selection is based on dispersion index, AIC, BIC and log likelihood statistics, putting into consideration Poisson, Negative Binomial, Zero Inflated Poisson and Zero Inflated Negative Binomial models and results obtained indicates that ZINB is the best models for analyzing count data in the presence of outliers and/or excess zero.

Keywords: Count data, Overdispersion, Excess zero, outliers, Goodness of fit, Poisson, Negative Binomial and Zero inflated models

1. Introduction

Not all overdispersion is real; apparent overdispersion can sometimes be identified and the model amended to eliminate it. Apparent overdispersion occurs when we can externally adjust the model to reduce the dispersion statistic closer to 1.0. It may occur because of a missing explanatory/predictor variable(s), the data contain outliers, the model requires an interaction term, a predictor needs to be transformed to another scale, or the link function is misspecified (Hardin and Hilbe 2007). When a real overdispersion in a model has been determined; then we employed another count model which can accommodate this problem.

For count data models considerable emphasis has been placed on analysis based on the assumption of correct specification of the conditional mean, or on the assumption of correct specification of both the conditional mean and conditional variance. i.e. $E[y|x]=\exp(X'\beta)$ $V[y|x]=\exp(X'\beta)$ since $E[y|x]=V[y|x]$

This is a nonlinear generalization of the linear regression model. It is a special case of the class of generalized linear models, widely used in statistics literature. Estimators for generalized linear models (GLMs) coincide with maximum likelihood estimators if the specified density is in the linear exponential family (Cameron and Trivedi 2008). The purpose of GLMs, and the linear models that they generalize, is to specify the relationship between the observed response variable and some number of covariates. The outcome variable is viewed as a realization from a random variable.

The study was aimed at examining the performance of some count models and how adequately did each model fit the data, base on dispersion indices, AIC, BIC and Log likelihood statistics. It further checks the biasness of each model in estimating the coefficient of the predictors used for the simulation.

2. Methodology

Impact of outliers and excess zero on count data were both studied, by creating simulated data set for 20, 50 and 100 sample sizes. Outliers were introduced into the generated data adding 5 to 5%, 10% and 15%, respective observation of y_i in the different data set generated, which have been randomized and replicated 500 times each for the respective selected sample sizes. Each constructed data set entails a specific cause of the overdispersion observed in the display of model output. We first create a base Poisson data set consisting of three normally distributed predictors as follows. Constant = 1, β_1 = 0.3, β_2 = -0.6, and β_3 = 0.4 which are coefficients of the predictors for sample size 20, 50 and 100. Poisson, Negative Binomial, Zero Inflated Poisson and Zero Inflated

Negative Binomial were considered to test how well each of the model fits the selected data sets having outliers and excess zero. The models were compared based on dispersion index in order to examine the changes made in the index when employed on the same set of data.

The basic count model is the Poisson regression model which is based on the Poisson distribution with probability density function:

$$\Pr(\lambda_i, y_i) = \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}. \qquad for\ y_i = 0,1,2,\ldots \tag{1}$$

where it is assumed that: $E(Y_i) = \lambda_i$; and $\lambda_i = exp(x\beta) = exp(1 + 0.3x_{1i} - 0.6x_{2i} + 0.4x_{3i})$ $Var(Y_i) = \lambda_i$. Thus, for the Poisson models, $E(Y_i) = Var(Y_i)$. The restrictive condition that the mean must equal the variance is often violated by overdispersed data (where variance exceeds the mean). As a result of that Poisson model is generally considered inappropriate for count data, which are usually highly skewed and overdispersed (Cameron and Trivedi 2008).

And the Negative binomial distribution function is given as follows;

$$\Pr(\mu_i, \alpha, y_i) = \frac{\Gamma\left(y_i + \frac{1}{\alpha}\right)}{\Gamma(y_i+1)\Gamma\left(\frac{1}{\alpha}\right)}\left(\frac{1}{1+\alpha\lambda_i}\right)^{1/\alpha}\left(\frac{\alpha\lambda_i}{1+\alpha\lambda_i}\right)^{y_i}, \qquad y_i = 0,1,2,\ldots \tag{2}$$

Here, the dispersion parameter $\alpha > 0$ and $\lambda_i = E(Y_i)$; and $Var(Y_i) = \lambda_i + \alpha\lambda_i^2$. The Negative Binomial model offers a practical solution to the overdispersion problem. However it does not address the issue of excess zeros (Wang 2007). Lawal (2011) argued that the Negative Binomial (NB) model might be a suitable alternative to the Poisson model especially for overdispersed count data. This is because the NB model in this case would account for the heterogeneity in the data by introducing the dispersion parameter α. The NB model (2) is equivalent to the Poisson model (1) when α equals zero. The larger the value of α is, the more variability in the data. The advantage of the NB model over the Poisson model can therefore be assessed by the significance of the α parameter (Lawal 2010).

Zero Inflated Poisson ZIP model has been considered by Lambert (1992) as a mixture of a zero point mass and a Poisson, while Heilborn (1989) similarly considers the Negative Binomial model case. Generally, for the Zero Inflated models, the probability of observing a zero outcome equals the probability that an observation is in the always zero group plus the probability that the observation is not in that group times the probability that the counting process produces a zero; Hilbe and Greene (2007). Therefore, the zero inflated probability mass function has the form:

$$\Pr(y_i) = \begin{cases} \psi + (1-\psi)Pr(Y=0) & if\ y_i = 0 \\ (1-\psi)Pr(Y = y_i) & if\ y_i > 0 \end{cases} \tag{3}$$

For the ZIP therefore, the probability mass function has:

$$Pr(\mu_i, y_i) = \begin{cases} \psi + (1-\psi)e^{-\lambda_i} & if\ y_i = 0 \\ (1-\psi)\frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!} & if\ y_i = 1,2,\ldots \end{cases} \tag{4}$$

such that $0 \le \psi < 1$. Thus the above model incorporates extra zeros than the original Poisson models in (1) in which $(\psi = 0)$. The mean and variance are respectively: $E(Y_i) = \lambda_i(1-\psi)$ and $Var(Y_i) = \lambda_i(1-\psi)(1 + \psi\lambda_i)$

The probability density function for a Zero Inflated Negative Binomial distribution ZINB is given by:

$$Pr(Y_i = y_i) = \begin{cases} \psi + (1-\psi)(1+\alpha\lambda_i)^{-\alpha^{-1}} & if\ y_i = 0 \\ (1-\psi)\frac{\Gamma(y_i+\alpha^{-1})}{y_i!\Gamma(\alpha^{-1})}\frac{(\alpha\lambda_i)^{y_i}}{(1+\alpha\lambda_i)^{y_i+\alpha^{-1}}} & if\ y_i > 0 \end{cases} \tag{5}$$

with $E(Y_i) = \lambda_i(1-\psi)$; and $Var(Y_i) = \lambda_i(1-\psi)(1 + \alpha\lambda_i + \psi\lambda_i)$ where the parameter λ_i and ψ depend on the covariates and $\alpha \ge 0$ is a scalar. Thus we have overdispersion whenever either ψ or α is greater than 0. The equation above reduces to Negative Binomial model (2) when $\psi = 0$ and to the ZIP when $\alpha = 0$.

The criteria for the assessment of the dispersion index was based on the criterion given by Hilbe (2008); if the dispersion index is greater than 1.0 the model may be overdispersed, if it is greater than 1.25 for models with moderate number of observations, then the model is overdispersed and if it is equal to or greater than 1.05 for models with large number of observation the model is also overdispersed. Log-likelihood as well as AIC and

BIC were computed for each model. The log-likelihood values were computed due to observation y_i for all the count models. For the Poisson model, the component of the log likelihood function for y_i is given by:

$$l_i = \lambda_i + y_i log(\lambda_i) - log(y_i!) \tag{6}$$

and of course, the log likelihood function for the Poisson is the sum of these terms over a random sample of size n,

$$l = -\sum_{i=1}^{n} \lambda_i + \sum_{i=1}^{n} y_i log(\lambda_i) - \sum_{i=1}^{n} log(y_i!) \tag{7}$$

Restricting ourselves to the component of the log likelihood functions therefore, we similarly have the log likelihood function component for y_i, having the Negative Binomial (NB) distribution as

$$l_i = log\Gamma\left(y_i + \frac{1}{\alpha}\right) - log\Gamma(y_i + 1) - log\Gamma\left(\frac{1}{\alpha}\right) + y_i log(\alpha\lambda_i) - \left(y_i + \frac{1}{\alpha}\right) log(1 + \alpha\lambda_i) \tag{8}$$

Considering the following indicator variables, $\omega 1$ and $\omega 2$ where $\omega 1$ equals 1 when observed count is zero and zero elsewhere. Similarly, $\omega 2$ equals 1 when observed counts are ≥ 1 and zero elsewhere. The use of these indicators ensures that the maximization of the log likelihood functions are uniform across the entire sample (Lawal 2010). Thus the indicator variables are defined as:

$$\omega 1 \begin{cases} 1 & if\ y_i = 0 \\ 0 & elsewhere \end{cases} \quad \text{and} \quad \omega 2 \begin{cases} 0 & if\ y_i = 0 \\ 1 & elsewhere \end{cases}$$

Consequently, the log likelihood functions for a given observation y_i are estimated as follows for the ZIP and ZINB models respectively in expressions in

$$l_i = \omega 1 \times \left[log\left(\psi + (1 - \psi)exp(\lambda_i)\right)\right] + \omega 2 \times [log(1 - \psi) + y_i log(\lambda_i) - log(y_i!) - \lambda_i] \tag{9}$$

$$l_i = \omega 1 \times \left[log\left(\psi + (1 - \psi)(1 + \alpha\lambda_i)^{-\alpha^{-1}}\right)\right] + \omega 2 \times [log(1 - \psi) + y_i log(\lambda_i) + y_i log\alpha - log(y_i!) - (y_i + \alpha^{-1})\log(1 + \alpha\lambda_i) + log\Gamma(y_i + \alpha^{-1}) - log\Gamma(\alpha^{-1})] \tag{10}$$

And the AIC and BIC were defined respectively as

$$AIC = -2lnL + 2k \tag{11}$$

$$BIC = -2lnL + kln(n) \tag{12}$$

where *lnL* is the overall likelihood and *k* is the number of parameters of the model. These formulae are from Akaike (1974) and Schwarz (1978) respectively. The criterion for the goodness of fit base on AIC and BIC is such that the lower the value of the statistic, the better fitting the model. While for log-likelihood the higher the value of the statistics the better fitting the model.

3. Discussion of Results

Result for the analysis of the four count models considered for this study were presented in Table 1- 10 below. Table 1 present the dispersion indices for the four models considered. Data set were analyzed at different Magnitude of outliers for three different sample sizes 20, 50 and 100. The dispersion indices of Negative Binomial at 0% magnitude of outliers are closer to 1 and considered the base for sample size 20 and 50, while for sample size 100, it has the same values with Poisson models. Meanwhile, Zero Inflated Negative Binomial has the least dispersion indices that are closer to 1 at 4%/5%, 10% and 15%/16% magnitude of outliers and considered the best models for all the sample sizes used.

Table 1: Effect of Outliers on Dispersion Indices for some Models for Count Data

Sample Size	Magnitude of Outlier(s)	Poisson	Negative Binomial	Zero Inflated Poisson	Zero Inflated Negative Binomial
20	0%	0.909625	**0.909627**	0.85612	0.831659
	5%	1.299387	1.124377	1.058237	**1.028002**
	10%	1.815388	1.173248	1.104233	**1.042887**
	15%	2.220315	1.198061	1.127587	**1.064943**
50	0%	0.813319	**0.813323**	0.796018	0.779435
	4%	1.046438	1.046437	1.024172	**1.002835**
	10%	1.265147	1.126912	1.079957	**1.068824**
	16%	1.488154	1.115822	1.092081	**1.069329**
100	0%	**0.891563**	**0.891563**	0.882372	0.873368
	5%	1.060416	1.060418	1.049486	**1.038777**
	10%	1.377747	1.249836	1.182111	**1.159268**
	15%	1.529250	1.195671	1.183344	**1.17127**

It still shows that Zero Inflated Negative Binomial fits the data more adequately than the other models used. Sample size 20 and 50 show absence of overdispersion based on the criteria given by (Hilbe 2008). This indicates that in the presence of outliers which cause overdispersion, Zero Inflated Negative Binomial is an alternative model for analyzing the count data and it fits the data more adequately than the other models used.

Table 2: Effect of Outliers on AIC Values for some Models for Count Data

Sample Size	Magnitude of Outlier(s)	Poisson	Negative Binomial	Zero Inflated Poisson	Zero Inflated Negative Binomial
20	0%	65.5987413	65.59874	**57.88864**	**57.88864**
	5%	72.1300147	71.91756	**63.11032**	**63.11032**
	10%	81.0298132	78.89772	**67.0719**	67.07032
	15%	88.1388642	83.71242	75.47814	**74.24234**
50	0%	184.898991	184.899	**175.2437**	**175.2437**
	4%	195.415266	195.4153	**185.1859**	**185.1859**
	10%	208.714849	208.3441	197.5757	**197.4791**
	16%	221.906446	219.4541	209.3426	**207.8082**
100	0%	362.385247	362.3853	**352.786**	**352.786**
	5%	380.89971	380.8997	**368.9704**	**368.9704**
	10%	412.16201	411.4715	401.5944	**401.5042**
	15%	432.420494	428.0277	419.9284	**417.9078**

Table 2 and 3 present the results for AIC and BIC, when five is added to certain percentages of the datasets for the four models. It can be seen clearly at 0% and 4%/5% magnitude of outliers, Zero Inflated Negative Binomial and Zero Inflated Poisson models outperform the other two models having the smallest AIC and BIC values and are considered the best. At 10% and 15%/16% magnitude of outliers Zero Inflated Negative Binomial model has the least values for AIC and BIC which was considered as the best model for all the sample size. Therefore, ZINB fits the data adequately well followed by ZIP for the selected samples, this shows that as the sample size increases, the best models for analyzing count data in presence of outlier is ZINB.

Table 3: Effect of Outliers on BIC Values for some Models for Count Data

Sample Size	Magnitude of Outlier(s)	Poisson	Negative Binomial	Zero Inflated Poisson	Zero Inflated Negative Binomial
20	0%	69.590206	69.59021	**61.8801**	**61.8801**
	5%	76.121479	75.90902	**67.10178**	**67.10178**
	10%	85.021278	82.88918	71.06336	**71.06178**
	15%	92.130329	87.70388	79.4696	**78.2338**
50	0%	190.72304	190.723	**181.0677**	**181.0677**
	4%	201.23931	201.2393	**191.01**	**191.01**
	10%	214.5389	214.1681	203.3997	**203.3032**
	16%	227.73049	225.2781	215.1666	**213.6322**
100	0%	369.59559	369.5956	**359.9963**	**359.9963**
	5%	388.11005	388.1101	**376.1807**	**376.1807**
	10%	419.37235	418.6818	408.8047	**408.7145**
	15%	439.63083	435.238	427.1387	**425.1181**

Table 4 present the log-likelihood values for the four count models, where ZINB has the least values at 10% and 15%/16% magnitude of outliers for all sample size. It has the same values with ZIP at 0% and 4%/5% magnitude of outliers for all the samples used for the study.

Table 4: Effect of Outliers on Log Likelihood Values for some Models for Count Data

Sample Size	Magnitude of Outlier(s)	Poisson	Negative Binomial	Zero Inflated Poisson	Zero Inflated Negative Binomial
20	0%	-31.79937065	-31.79937165	**-27.94432**	**-27.94432**
	5%	-35.06500733	-34.95878022	**-30.55516**	**-30.55516**
	10%	-39.51490659	-38.44885778	-32.53595	**-32.53516**
	15%	-43.0694321	-40.85620988	-36.73907	**-36.12117**
50	0%	-91.4494953	-91.44949979	**-86.62183**	**-86.62183**
	4%	-96.70763324	-96.70763469	**-91.59297**	**-91.59297**
	10%	-103.3574247	-103.172025	-97.78785	**-97.73957**
	16%	-109.953223	-108.7270507	-103.6713	**-102.9041**
100	0%	-180.1926233	-180.1926276	**-175.393**	**-175.393**
	5%	-189.4498551	-189.4498584	**-183.4852**	**-183.4852**
	10%	-205.0810048	-204.7357409	-199.7972	**-199.7521**
	15%	-215.2102472	-213.0138257	-208.9642	**-207.9539**

This study shows that the ZINB outperform the other models base on log-likelihood values followed by ZIP, this implies that ZINB can fit the data well in the presence of outlier even if the sample size continue to be increased.

Table 5 show the bias in each of the count models in estimating the parameters used to simulate the data set. Based on average performance Negative Binomial has the least bias at most of the percentages of magnitude of outliers for sample size 20 and 100, while it has almost the same performance with Poisson model for sample size 50.

Table 5: Bias of Coefficient of the Predictors and the Constant Term

Parameters	Sample Size	Magnitude of Outlier(s)	Poisson	Negative Binomial	Zero Inflated Poisson	Zero Inflated Negative Binomial
$\beta_1=0.3$	20	0%	**0.017873**	0.017871	-0.06074	-0.06076
$\beta_1=0.3$	20	5%	-0.02795	**-0.01341**	-0.111	-0.11102
$\beta_1=0.3$	20	10%	-0.06709	**-0.01891**	-0.1998	-0.19801
$\beta_1=0.3$	20	15%	0.155222	0.208619	**0.05076**	0.083411
$\beta_2=-0.6$	20	0%	**-0.03639**	**-0.03639**	-0.16631	-0.1663
$\beta_2=-0.6$	20	5%	-0.04697	**-0.039**	-0.18203	-0.18203
$\beta_2=-0.6$	20	10%	-0.03816	**0.000261**	-0.20185	-0.20231
$\beta_2=-0.6$	20	15%	-0.0935	**-0.00832**	-0.23989	-0.22971
$\beta_3=0.4$	20	0%	0.002272	**0.002271**	0.212897	0.212861
$\beta_3=0.4$	20	5%	0.007962	**-0.00021**	0.229297	0.229287
$\beta_3=0.4$	20	10%	0.284976	**0.231656**	0.516835	0.514075
$\beta_3=0.4$	20	15%	0.191124	**0.14423**	0.387168	0.353873
$\mu=1.0$	20	0%	**0.014217**	0.014219	-0.12349	-0.12346
$\mu=1.0$	20	5%	**-0.09129**	-0.09149	-0.23177	-0.23176
$\mu=1.0$	20	10%	**-0.25887**	-0.26448	-0.39607	-0.39515
$\mu=1.0$	20	15%	**-0.30881**	-0.33022	-0.44096	-0.42993
$\beta_1=0.3$	50	0%	**0.056844**	0.056843	0.101167	0.101167
$\beta_1=0.3$	50	4%	**0.051205**	**0.051205**	0.095763	0.095766
$\beta_1=0.3$	50	10%	0.071222	**0.070366**	0.115094	0.115775
$\beta_1=0.3$	50	16%	**0.082109**	0.085782	0.127113	0.13355
$\beta_2=-0.6$	50	0%	-0.06835	**-0.06834**	-0.13581	-0.1358
$\beta_2=-0.6$	50	4%	**-0.07213**	**-0.07213**	-0.13986	-0.13989
$\beta_2=-0.6$	50	10%	-0.0652	**-0.05789**	-0.13276	-0.13127
$\beta_2=-0.6$	50	16%	-0.08492	**-0.06508**	-0.15477	-0.14818
$\beta_3=0.4$	50	0%	-0.02737	-0.02738	**0.015981**	0.015982
$\beta_3=0.4$	50	4%	**-0.02012**	**-0.02012**	0.023426	0.023434
$\beta_3=0.4$	50	10%	**0.02493**	0.02816	0.06816	0.070359
$\beta_3=0.4$	50	16%	**0.041065**	0.053538	0.085539	0.096626
$\mu=1.0$	50	0%	-0.12725	**-0.12724**	-0.19667	-0.19668
$\mu=1.0$	50	4%	**-0.18213**	**-0.18213**	-0.25174	-0.25173
$\mu=1.0$	50	10%	-0.26467	**-0.26343**	-0.33403	-0.33395
$\mu=1.0$	50	16%	-0.34665	**-0.34464**	-0.41734	-0.41767
$\beta_1=0.3$	100	0%	**0.018751**	**0.018751**	0.05136	0.051297
$\beta_1=0.3$	100	5%	0.075697	**0.075696**	0.111532	0.111516
$\beta_1=0.3$	100	10%	0.059218	**0.05525**	0.117686	0.116046
$\beta_1=0.3$	100	15%	0.072671	**0.059415**	0.136144	0.126337
$\beta_2=-0.6$	100	0%	**-0.03727**	**-0.03727**	-0.07068	-0.07065
$\beta_2=-0.6$	100	5%	**-0.04403**	**-0.04403**	-0.08119	-0.08118
$\beta_2=-0.6$	100	10%	-0.11215	**-0.10744**	-0.16952	-0.16711
$\beta_2=-0.6$	100	15%	-0.09346	**-0.07722**	-0.15911	-0.14374
$\beta_3=0.4$	100	0%	**0.066833**	**0.066833**	0.081928	0.081936
$\beta_3=0.4$	100	5%	**0.078727**	**0.078727**	0.095763	0.095768
$\beta_3=0.4$	100	10%	0.123232	**0.116181**	0.147047	0.144468
$\beta_3=0.4$	100	15%	0.148301	**0.130668**	0.175631	0.163087
$\mu=1.0$	100	0%	**-0.01067**	**-0.01067**	-0.05167	-0.05175
$\mu=1.0$	100	5%	**-0.09753**	**-0.09753**	-0.14152	-0.14151
$\mu=1.0$	100	10%	-0.20383	**-0.20173**	-0.27697	-0.27396
$\mu=1.0$	100	15%	-0.26427	**-0.25782**	-0.34798	-0.33227

In terms of biasness, Negative Binomial outperforms the other models, even though there are no significant differences in the comparison base on the magnitude of outliers present in the data set. This can also be seen clearly in Table 5 that shows the bias of each of the models in estimating the model parameters.

The four count models were also employed on the samples incorporated with excess zero and the results based on dispersion indices, AIC, BIC and log likelihood were presented in Tables 6 - 10 respectively. And it can be notice that in Table 6 when Negative Binomial was employed the dispersion indices decreases slightly close to 1, it further dropped closer to 1 when ZIP and ZINB were employed. From the result of this Table Negative Binomial has the closest dispersion indices to 1 at constant 1.0 and considered the best for sample size 20 and 50, while Poisson model has the least dispersion indices closer to 1 and considered the best for the sample size 100 at constant 1.0. ZIP outperform other models having dispersion indices closest to 1 for sample size 20 and 50 at constant 0.5 and considered the best model. For sample size 100 ZINB outperform other models at constant 0.5 and 0.2 and still considered the best at constant 0.2 for sample size 20 and 50.

Table 6: Effect of Excess Zero on Dispersion Indices for some Models for Count Data

Sample Size	Constant	Poisson	Negative Binomial	Zero Inflated Poisson	Zero Inflated Negative Binomial
20	1.0	.9096255	.9096272	0.8561197	**0.808558**
	0.5	1.038108	1.038107	0.9770419	**0.922762**
	0.2	1.246094	1.246091	1.1727915	**1.107636**
50	1.0	.8133193	.8133235	0.7960187	**0.779435**
	0.5	1.013668	1.013668	0.9921006	**0.971432**
	0.2	1.418688	1.247437	1.1710633	**1.103502**
100	1.0	.8915636	.8915633	0.8823719	**0.873368**
	0.5	1.360607	1.21868	1.1699328	**1.083271**
	0.2	1.196868	1.132696	1.0873882	**1.035608**

One can also notice from these results that the overdispersion was taking care base on the underlying criteria by (Hilbe 2008). AIC values were presented in Table 7 where the values decrease with increase in magnitude of zeros for the four selected count models. ZINB and ZIP has the least AIC values for sample size 20 and 100 at constant 1.0, 0.5 and 0.2 and considered the best model. While for sample size 50, the two models have the same least AIC values at constant 1.0 and 0.5, but ZINB outperform the other models at constant 0.2 and considered the best. This implies that ZINB fits the data more adequately than other models.

Table 7: Effect of Excess Zero on AIC Values for some Models for Count Data

Sample Size	Constant	Poisson	Negative Binomial	Zero Inflated Poisson	Zero Inflated Negative Binomial
20	1.0	65.5987413	65.59874	**57.88864**	**57.88864**
	0.5	60.0285349	60.02853	**54.90542**	**54.90542**
	0.2	54.0298936	54.0299	**48.28804**	**48.28804**
50	1.0	184.898991	184.899	**175.2437**	**175.2437**
	0.5	163.8083	163.8083	**163.6985**	**163.6985**
	0.2	158.012365	157.5883	156.9974	**156.1711**
100	1.0	362.385247	362.3853	**352.786**	**352.786**
	0.5	325.306066	324.6847	**319.824**	**319.824**
	0.2	269.988107	269.8002	**262.3872**	**262.3872**

The BIC values presented in Table 8 decreased with increase in magnitude of zeros for the four count models used for this study. ZINB and ZIP have the least BIC values for sample size 20 and 100 at constant 1.0, 0.5 and 0.2 and considered the best model. While for sample size 50, the two models have the same least BIC values at constant 1.0 and 0.5, but ZINB outperform the other models at constant 0.2 and considered the best. This indicates that ZINB is the best models for analyzing count data having excess zero.

Table 8: Effect of Excess Zero on BIC Values for some Models for Count Data

Sample Size	Constant	Poisson	Negative Binomial	Zero Inflated Poisson	Zero Inflated Negative Binomial
20	1.0	69.590206	69.59021	**61.8801**	**61.8801**
	0.5	64.019999	64.0200	**58.89688**	**58.89688**
	0.2	58.021358	58.02136	**52.2795**	**52.2795**
50	1.0	190.72304	190.723	**181.0677**	**181.0677**
	0.5	169.63235	169.6323	**169.5225**	**169.5225**
	0.2	163.83641	163.4124	162.8215	**161.9951**
100	1.0	369.59559	369.5956	**359.9963**	**359.9963**
	0.5	332.51641	331.895	**327.0343**	**327.0343**
	0.2	277.19845	277.0105	**269.5975**	**269.5975**

Table 9 shows the log likelihood values for the four count models considered. ZIP and ZINB have the highest log likelihood values at constant 1.0, 0.5 and 0.2 for sample size 20 and 100 which were considered the best models.

Table 9: Effect of Excess Zero on Log likelihood values for some Models for Count Data

Sample Size	Constant	Poisson	Negative Binomial	Zero Inflated Poisson	Zero Inflated Negative Binomial
20	1.0	-31.79937065	-31.79937165	**-27.94432**	**-27.94432**
	0.5	-29.01426744	-29.01426735	**-26.45271**	**-26.45271**
	0.2	-26.0149468	-26.0149481	**-23.14402**	**-23.14402**
50	1.0	-91.4494953	-91.44949979	**-86.62183**	**-86.62183**
	0.5	-80.90415002	-80.90415129	**-80.84925**	**-80.84925**
	0.2	-78.00618245	-77.79416546	-77.49872	**-77.08555**
100	1.0	-180.1926233	-180.1926276	**-175.393**	**-175.393**
	0.5	-161.653033	-161.3423372	**-158.912**	**-158.912**
	0.2	-133.9940535	-133.9000798	**-130.1936**	**-130.1936**

The ZINB also has the highest log-likelihood values for sample size 50 at constant 0.2. Meanwhile it maintains the same highest values with ZIP at constant 1.0 and 0.5 and considered the best models for the same sample size 50. This indicates that the ZINB outperform the other models in analyzing count data with excess zero.

Table 10: Bias of Coefficients of the Predictors and the Constant Term (Excess Zero)

Parameters	Sample Size	Constant	Poisson	Negative Binomial	Zero Inflated Poisson	Zero Inflated Negative Binomial
$\beta_1=0.3$	20	1.0	0.017873	**0.017871**	-0.06074	-0.06076
$\beta_1=0.3$	20	0.5	-0.07283	-0.07283	0.043609	**0.043603**
$\beta_1=0.3$	20	0.2	**-0.01202**	**-0.01202**	0.123124	0.123112
$\beta_2=-0.6$	20	1.0	**-0.03639**	**-0.03639**	-0.16631	-0.1663
$\beta_2=-0.6$	20	0.5	0.154509	0.15451	-0.00121	**-0.00115**
$\beta_2=-0.6$	20	0.2	**0.001201**	0.001206	-0.12533	-0.12534
$\beta_3=0.4$	20	1.0	0.002272	**0.002271**	0.212897	0.212861
$\beta_3=0.4$	20	0.5	0.066403	0.066403	0.045691	**0.045676**
$\beta_3=0.4$	20	0.2	**-0.02463**	**-0.02463**	0.049864	0.049781
$\mu=1.0$	20	1.0	**0.014217**	0.014219	-0.12349	-0.12346
$\mu=0.5$	20	0.5	0.3278	0.3278	**0.054334**	0.054409
$\mu=0.2$	20	0.2	**-0.02497**	**-0.02497**	-0.25094	-0.25087
$\beta_1=0.3$	50	1.0	0.056844	**0.056843**	0.101167	0.101167
$\beta_1=0.3$	50	0.5	-0.0721	-0.0721	-0.05934	**-0.03097**
$\beta_1=0.3$	50	0.2	-0.01228	**-0.00852**	-0.12168	0.2177
$\beta_2=-0.6$	50	1.0	-0.06835	**-0.06834**	-0.13581	-0.1358
$\beta_2=-0.6$	50	0.5	**-0.02612**	**-0.02612**	-0.03486	-0.03486
$\beta_2=-0.6$	50	0.2	0.042066	**0.03812**	-0.11776	-0.5072
$\beta_3=0.4$	50	1.0	-0.02737	-0.02738	**0.015981**	0.015982
$\beta_3=0.4$	50	0.5	**-0.00746**	**-0.00746**	0.007974	0.007968
$\beta_3=0.4$	50	0.2	**-0.00349**	0.005735	0.202991	0.371622
$\mu=1.0$	50	1.0	-0.12725	**-0.12724**	-0.19667	-0.19668
$\mu=0.5$	50	0.5	**-0.01429**	**-0.01429**	-0.04531	-0.0453
$\mu=0.2$	50	0.2	0.004478	**0.001025**	-0.17213	-0.39213
$\beta_1=0.3$	100	1.0	**0.018751**	**0.018751**	0.05136	0.051297
$\beta_1=0.3$	100	0.5	**-0.02259**	-0.02588	0.035889	0.035891
$\beta_1=0.3$	100	0.2	-0.0155	**-0.01542**	-0.03097	-0.0213
$\beta_2=-0.6$	100	1.0	**-0.03727**	**-0.03727**	-0.07068	-0.07065
$\beta_2=-0.6$	100	0.5	-0.06216	-0.07164	-0.01684	**-0.01678**
$\beta_2=-0.6$	100	0.2	0.129651	0.124326	**0.074513**	0.074523
$\beta_3=0.4$	100	1.0	**0.066833**	**0.066833**	0.081928	0.081936
$\beta_3=0.4$	100	0.5	-0.03406	-0.02757	**-0.00941**	-0.00945
$\beta_3=0.4$	100	0.2	0.036406	**0.033128**	-0.04842	-0.04841
$\mu=1.0$	100	1.0	**-0.01067**	**-0.01067**	-0.05167	-0.05175
$\mu=0.5$	100	0.5	0.064403	0.061059	-0.01009	**-0.01002**
$\mu=0.2$	100	0.2	0.122264	0.120995	**0.042923**	0.04294

Table 10 shows the biasness of the four count models in terms of estimating the parameters used in simulating the dataset. Negative Binomial model outperform other models in estimating some of the model parameters for sample sizes used followed by Poisson model. However ZINB and ZIP models were considered the best in estimating few parameters for sample size 20 and 100. The amount of biasness fluctuates with increase in the magnitude of excess zero for all the sample size.

4. Conclusion

In conclusion, we found from our study that, the dispersion indices increases with increase in Magnitude of outliers and excess zero in the datasets considered for all the sample sizes. When ZINB was employed on the same data set the indices dropped closer to 1 which indicates that the model fits the data more adequately than the other models in terms of accommodating the problem of overdispersion. However, when sample size increased the dispersion indices decreases in all Magnitude of outliers. The AIC and BIC statistics of ZINB were the least for the analysis, followed by ZIP and Negative Binomial respectively. The statistics seriously increases as the sample size increased. While the AIC and BIC values decreases with increase in the magnitude of excess zeros. ZINB has the highest values of log likelihood statistics and the statistics increases with increase in sample sizes. Negative Binomial and Poisson models fit the data wells in terms of biasness for some parameters, meanwhile ZINB outperform other models in the remaining parameters. The amount of biasness decreases with increase in the magnitude of excess zero. These indicate that ZINB is the best models for analyzing count data in the presence of outliers and/or excess zero.

This study despite that it is time consuming, but if applied appropriately it will assist researchers to understand what proportion of outliers or excess zero may cause serious overdispersion to their work. This may give them a kind of an overview of their study. Most real life count data exhibit excessive zero, therefore, going by this study as well Zero Inflated Negative Binomial model can be use to analyze any count data more especially if its distribution pattern cannot be identify. Finally, the study can also be extended further on very large sample to investigate the performance of these models in the presence of outliers and/or excess zero.

Reference

Akaike, H. (1974), “A new look at the statistical model identification”, *IEEE Transactions on Automatic Control* 19: 716-723.

Cameron, C. A. and Trivedi, P. K.(2008), “Regression Analysis of Count Data”, Econometric Society Monographs No. 30, Reprinted edition, Cambridge University Press, New York, pp 1, 96-97, 221

Hardin, J. W. and Hilbe, J. M. (2007), “Generalized Linear Models and Extensions”, Second Edition, A Stata Press Publication, StataCorp LP, College Station, Texas, pp 221

Heilborn, D. (1989), “Generalized Linear Models for Altered Zero Probability in Count Data”, Technical Report, Department of Epidemiology and Biostatistics, University of California, San Francisco.

Hilbe, J. M. (2008), “Negative Binomial Regression”, Cambridge University Press. New York. Copyright. Pp 51-64.

Hilbe, J. M. and W. Green (2007), “Count Response Regression Models”, In Epidemiology and Medical Statistics, Elsevier Handbook of Statistics series, ed. C. Rao, J. Miller, and D. Rao. London: Elsevier.

Johansson, P., (1996), “Speed Limitation and Motorway Casualties: A Time Series Count Data Regression Approach”, *Accident Analysis and Prevention* 28, 73-87.

Lawal, H. B. (2010), “Zero-Inflated Count Regression Models with Application to Some Examples”, Published online: 3 April 2010.© Springer Science+Business Media B.V. 2010.pp 21

Lawal, H. B. (2011), “On Zero-Truncated Generalized Poisson Count Regression Model”, *Pioneer Journal of Theoretical and Applied Statistics*, Vol 1, Number 1, 2011, pp 16

Lambert, D. (1992), “Zero Inflated Poisson Regression with an Application to Defects in Manufacturing”, *Technometric* vol. 34, 3

Wang, R., (2007), “A Zero Augmented Negative Binomial Model in the Analysis of Medical Care Count Data”, Department of Economic, University of Hawaii at Manoa.

Biography: Mohammed Usman (NSA member’07). He was born in Gwoza, Gwoza Local Govt, Borno State Nigeria in June 1971 and has been at the Federal Polytechnic Bali, Taraba State Nigeria since 2010 as a pioneer Head of Department of Statistics, he worked with Federal Polytechnic Mubi, Math and Statistics department between 2001-2010, and Umar Ibn Ibrahim Elkanemi College of Education Science and Technology between 2000-2001. He attended Gwange IV Primary School Maiduguri between 1979-1984, Government Teachers College Gwoza between 1984-1989, University of Maiduguri between 1991-1997 and University of Ilorin, 2006-2007 and 2008 to date. He holds Diploma Mathematics Education, a Bachalor and Masters Degree in Statistics and presently doing Doctorate Degree in Statistics. He is a member of the Nigerian Statistical Association (NSA), Teachers Registration Council of Nigeria, Nigerian Educational Research Association (NERA) and the International Research and Development Institute (IRDI) Network. He is presently working on Statistical modeling of count data in the presence of outliers and excess zero.