Original Research Article

COMPARING ZEROINFLATED POISSON, ZERO-INFLATED NEGATIVE BINOMIAL AND ZERO-INFLATED GEOMETRIC IN COUNT DATA WITH EXCESS ZERO

ABSTRACT

Count data often violate the assumptions of a normal distribution due to the fact that they are bounded by their lowest value which is zero. The Poison distribution is sometimes suggested but when the assumption of equal mean and variance is violated due to over-dispersion and presence of zeros we tend to look in the direction of other models. Zero-inflated data falls in this category. The zero-inflated and hurdle models have been found to fit this scenario. The proportions of zero in the data often affect the choice of the models. Our study used the Monte Carlo design to sample 1000 cases from positively skewed distribution with 1.25 as mean vector and 0.10 as zero-inflation parameter. The data was analysed using the method of the maximum likelihood estimation. The Zero-Inflated Poisson, Zero-Inflated Negative Binomial and Zero-Inflated Geometric were fitted; the standard error and Akaike Information Criterion were obtained as measures of model validation with ZIP outperformed ZINB and ZIG.

1.0 INTRODUCTION

In any statistical data analysis, it is necessary to determine the type of data being analysed. The data would assume to fulfill a basic assumption of normal distribution. Many other distributions equally exist. When the distribution assumed is at variance with the actual distribution of the data, the validity of the results will show in the dissimilarity between the data and the distribution assumed in the analysis. It is therefore imperative for researchers to choose a distribution similar to that which the data possesses. Counts data belongs to the

class of data which is at variance to the assumption of a normal distribution; this is because of the fact that counts data are bounded by their lowest value which is zero. Therefore, Poisson distribution with a log link is often been assumed and preferred over and against normal distribution with Gaussian link. The use of Poisson distribution may not guarantee valid results due to the fact that other features inherent in the data may invalidate Poisson assumption thereby paving the way for researchers to examine more accurate and valid model. Among such models are the two-part models such as the hurdle model and zero-inflated models. They are class of models that can handle data with excess zeros.

2. LITERATURE REVIREW

count data are constraint by their lower bound zero value, this however makes it difficult in analysing a count data because assumptions of normality become invalid because the data is either positively skewed or negatively skewed depending on the proportion of zero that is in the data. The data is heteroscedastic in nature with variance increasing as the count increases (Jeffrey 2007). Under this circumstance it is therefore, inappropriate to use model like ordinary least square regression which is strictly based on the assumption of normality. That is the residuals are distributed normally with a mean zero and standard deviation of one. Invariably Cameron and Triverdi (1998) have found the OLS regression to be suitable for count data only when the mean of the count is high.

However, in handling zero inflated data many solutions were prescribed in the literature. One of the simplest of solutions was to delete all cases having responses of zero on the variable of interest. A large proportion of total responses would then be removed from the total dataset. This method would result in loss of useful and valuable information and would have adverse consequence on statistical conclusion validity (Tooze, Grunwald, & Jones, 2002). However, the sample size may become too small for analyses. Another solution prescribed was to ignore the zero-inflation, assume asymptotic normality, and analyze the data using standard techniques such as ordinary least squares regression.

$$Y_i = \alpha + \beta_1 X_1 + \cdots + \beta_k X_k + \varepsilon_i \quad 1.0$$

The model assumes that the residuals for Y_i are distributed normally with a mean of zero and a common variance, σ^2. For the first equation, **y** is a vector of responses, **X** is a design matrix for the explanatory variable responses, β is a vector of regression coefficients relating **y** to **X**, and ε is a vector of residuals measuring the deviation between the observed values of the design matrix and those predicted from the fitted equation.

However, since we are dealing with nonlinear models, these phenomena belong to the class of generalized linear models (GLM) such as the Poisson Regression, Negative Binomial regression, ZIP, ZINB, ZIGP and hosts of others.

Poisson and negative binomial regression models are designed to analyze count data. The "rare events" nature of crime counts are controlled for in the formulas of both Poisson and negative binomial regression. However, Poisson and negative binomial regression models differ in regards to their assumptions of the conditional mean and variance of the dependent variable. Poisson models assume that the conditional mean and variance of the distribution are equal. Negative binomial regression models do not assume an equal mean and variance and particularly correct for over-dispersion in the data, which is when the variance is greater than the conditional mean (Osgood, 2000; Paternoster & Brame, 1997).

3. METHODOLOGY

Zero-Truncated Negative Binomial Regression Model

Let Yi be the nonnegative values random variable and suppose Y = 0 is observed with a frequency significantly higher than can be modeled by the usual Poisson model. Thus the regression model is defined by

$$P(Y_i = y_i/x_i, z_i) = \begin{cases} \varphi_i + (1-\varphi_i) f(0, \theta_i), & y_i = 0 \\ (1-\varphi_i) f(\ y_i, \theta_i), & y_i > 0 \end{cases} \quad 2.0$$

When $f(\ y_i, \theta_i)$, y_i = 0,1,2, … is the pdf of Y_i and $0 < \varphi_i < 1. The\ function\ \varphi_i = \varphi_i(Z_i)$ satisfies logit $(\varphi_i) = \log(\varphi_i(1-\varphi_i)^{-1}) = \sum_{j=1}^{m} Z_{ij}\delta_j$ where $Z_i = (Z_{i1}, Z_{i2}, \dots, Z_{im})$ the i^{th} row of the covariate matrix Z and $\delta = (\delta_1, \delta_2, \dots, \delta_m)$ are the unknown m-dimensional column vector of parameters. The nonnegative function $\varphi_i$is modeled via logit link function that allows φ_i being negative may be used.

We consider a zero-inflated negative binomial regression model in which a zero-inflated negative binomial regression model in which the response variable Y_i (i= 1, 2, …, n) has the distribution

$$\begin{cases} \varphi_i + (1-\varphi_i)\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu_i}\right)^{\alpha^{-1}}, y_i = 0 \\ (1-\varphi_i)\frac{\Gamma(y_i+\alpha^{-1})}{\Gamma(y_i+1)\,\Gamma(\alpha^{-1})}\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu_i}\right)^{\alpha^{-1}}\left(\frac{\mu_i}{\alpha^{-1}+\mu_i}\right)^{y_i}, y_i > 0 \end{cases} \quad \dots\dots\dots\dots\dots\dots 3.0$$

Where α(≥ 0)is a dispersion parameter that is assumed not to depend on covariates. Furthermore the model in 2 reduces to the ZIP distribution when parameter α → 0 and parameter $\lambda_i(X_i) and\ \varphi_i$satisfy

$\log(\lambda_i) = \sum_{j=1}^{m} X_{ij}\beta_j\ and\ o < \varphi_i < 1$. The mean and the variance of the distribution are E(Y_i) = $(1-\varphi_i)\mu_i$ and $\mathrm{Var(Yi)} = (1-\varphi_i)\mu_i(1+\varphi_i\mu_i+\alpha\mu_i)$.

Consider variable Y_i as a response variable which follows by a discrete distribution Pr(Y_i=y_i). For some observations, the value of Y_i may be truncated. If truncated for the i^{th} observation, we have $\mathrm{Yi} \geq \mathrm{y}_i$ (right truncation) and that observation is omitted to analyze from the data set. Thus the probability function for a right truncated variable Y_i can be written as

$$f_T = (y_i = \theta_i) = \frac{f(y_i,\theta_i)}{1-P(Y_i \geq y_i)}\ , \text{i=1, …, k} \qquad 4.0$$

When k is the number of observations after truncation, we can write the log likelihood of the truncation count regression model as

$$logL(\theta_i, Y_i) = \sum_{i=}^{k}(\log f(y_i, \theta_i) - \log(1 - \Pr(Y_i \geq y_i)) \quad 5.0$$

By taking partial derivatives with respect to θ and equal to zero we can obtain the parameter estimation. However, if we replace the function $f(y_i, \theta_i)$ into the negative binomial distribution model, the distribution with right truncation will be obtained as follow:

$$\Pr(Y_i = y_i) = \begin{cases} \frac{\varphi_i + (1-\varphi_i)\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu}\right)^{\alpha^{-1}}}{1-\sum_{y_{i+1}}^{\infty}(1-\varphi_i)g(y_i;\mu,\alpha)}, & y = 0 \\ \frac{(1-\varphi_i)g(y_i;\mu,\alpha)}{1-\sum_{y_{i+1}}^{\infty}(1-\varphi_i)g(y_i;\mu,\alpha)}, & 1 < y_i < t_\mu \end{cases} \quad 6.0$$

Where

$$g(y_i;\ \mu_i, \alpha) = \frac{\Gamma(y_i+\alpha^{-1})}{\Gamma(y_i+1)\,\Gamma(\alpha^{-1})}\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu_i}\right)^{\alpha^{-1}}\left(\frac{\mu_i}{\alpha^{-1}+\mu_i}\right)^{y_i}$$

And t_i is the truncation point for y_i which means that when $y_i > t_i$ we truncate the response variable. We can obtain the log-likelihood function for ZINB regression model with right truncations as follows:

$$LL_{(TZINB)} = \sum_{i=1}^{k}\begin{Bmatrix} I_{\{y_i=0}\log\left[\varphi_i + (1-\varphi_i)\left[\frac{\alpha^{-1}}{\alpha^{-1}+\mu_i}\right]^{\alpha^{-1}}\right] - log\left[1 - \sum_{y_i=t_i+1}^{\infty}(1-\Phi_i)g\right] \\ +I_{\{1\leq y_i\leq t_i}\left(\log(1-\varphi_i) + logg - \log\left[1 - \sum_{y_i=t_i+1}^{\infty}(1-\Phi_i)g\right]\right)\} \end{Bmatrix} \quad \ldots.7.0$$

Where k is the number of observation after truncation and the expression log $g(y_i; \mu_i, \alpha)$ can be obtained as follow:

$$\log g(y_i; \mu_i, \alpha) = \sum_{j=o}^{y_i-1}\log(j + \alpha^{-1}) - logy_i! + y_i log\alpha\mu_i - y_i\log(1 + \alpha\mu_i) \quad 8.0$$

The parameters estimation is obtained by MLE method. By taking the partial derivatives of the likelihood function and setting them to zero, the likelihood equations for estimating the parameters are obtained thus:

$$\frac{\partial LL_{(TZIN)}}{\partial \beta_r} = \sum_{i=1}^{\infty}\left\{ I_{(y_i=0)}\left[\frac{1-(1+\alpha\mu_i)^{1-\alpha^{-1}-1}}{W_i+(1+\alpha\mu_i)^{-\alpha^{-1}}} + \frac{\sum_{y=t_i+1}^{\infty}(1-\varphi_i)g\,\frac{y_i+2\alpha\mu_i}{\mu_i(1+\alpha\mu_i}}{1-\sum_{y_i=t_i+1}^{\infty}(1-\varphi_i)g}\right]x_{ir}\mu_i + \right.$$

$$\left. I_{1\le y\le t_i}\left[y_i\left(\frac{1}{\mu_i}-\frac{\alpha}{1+\alpha\mu_i}\right)+\frac{\sum_{y=t_i+1}^{\infty}(1-\varphi_i)g\frac{y_i+2\alpha\mu_i}{\mu_i(1+\alpha\mu_i}}{1-\sum_{y_i=t_i+1}^{\infty}(1-\varphi_i)g}x_{ir}\mu_i\right]\right\}\ldots\ldots \quad 9.0$$

$$\frac{\partial LL_{(TZIN)}}{\partial \alpha} = \sum_{I=1}^{\infty}\left\{ I_{\{Y_I=0}\left[\alpha^{\{-1\}\log(1+\alpha\mu_i)} - \frac{\mu_i}{1+\alpha\mu_i}\,\frac{1}{\alpha} + \left[\sum_{y_i=t_i+1}^{\infty}(1-\varphi_i)\left(\frac{\Gamma'(y_i+\alpha^{-1})}{\Gamma(y_i+\alpha^{-1})}-\frac{\Gamma'(\alpha^{-1})}{\Gamma(\alpha^{-1})}+\right.\right.\right.$$

$$\left.\left.\left.\alpha^{-2}\log(1+\alpha\mu_i)-\frac{\alpha^{-1}}{1+\alpha\mu_i}(\mu_i-y_i)\right)g\right] / \left(1-\sum_{y_i=t_i+1}^{\infty}(1-\varphi_i)g\right)+\right.$$

$$I_{\{1\le y_i\le t_i}\left[-\sum_{j=0}^{y_{i+1}}\frac{\alpha^{-2}}{j+\alpha^{-1}} - \frac{y_i\mu_i}{1+\alpha\,\mu_i}+\alpha^{\{-1\}y}\right]+\left[\sum_{y_i=t_i+1}^{\infty}(1-\varphi_i)\left(\frac{\Gamma'(y_i+\alpha^{-1})}{\Gamma(y_i+\alpha^{-1})}-\frac{\Gamma'(\alpha^{-1})}{\Gamma(\alpha^{-1})}+\right.\right.$$

$$\left.\left.\alpha^{-2}\log(1+\alpha\mu_i)-\frac{\alpha^{-1}}{1+\alpha\mu_i}(\mu_i-y_i)\right)g\right] / \left(1-\sum_{y_i=t_i+1}^{\infty}(1-\varphi_i)g\right) \;=\; 0$$

10.0

$$\frac{\partial LL_{(TZIN)}}{\partial \delta_s} =$$

$$\sum_{i=1}^{k}\left\{ I_{y_i=0}\left[\frac{[1-(1+\alpha\mu_i)^{-\alpha^{-1}}}{w_i+(1+\alpha\mu_i)^{-\alpha^{-1}}}-\frac{\sum_{y_i=t_i+1}^{\infty}g}{(1-\varphi_i)^{\{-1\}}-\sum_{y_i=t_i+1}^{\infty}g}\right]\varphi_i Z_{is} - I_{\{1\le y\le t_i\}}\left[\frac{\sum_{y_i=t_i+1}^{\infty}g}{(1-\varphi_i)^{\{-1\}}-\sum_{y_i=t_i+1}^{\infty}g}\right]\varphi_i Z_{is}\right\}$$

=0 11.0

4.0 DATA SIMULATION

The dependent variable y was simulated by the zero-inflation Poisson by setting the vector of mean as 1.25 and the zero-inflated parameter as 0.1. at n=15, proportion of zero was 0.2 with expected value equal 1.467 and variance equal 1.552. At n=25, p=0.32, expected value of 1.4 and variance equal1.75. At n=50, p=0.42, mean equal 1.14 and variance equal 1.551. At n=100, p=0.41, mean equal 0.96 and variance equal1.069. At n=150, p=0.33, mean equal

1.267 and variance equal 1.489. At n=300, p=0.377, mean equal equal 1.177 and variance equal 1.49. At n=500, p=0.366, mean =1.104, variance equal 1.199 and at n=1000, p=0.539, mean equal 1.77 variance equal 1.311

5.0 RESULT AND DISCUSSION

The results from the simulation showed that all the distributions were characterized with over dispersion with varying degrees of zero fractions.

Table 1a: ESTIMATES OF PARAMETERS AND THE AIC FOR THE COUNT PART

sample (n)	MODEL						AIC
15	(ZIP)	0.08246	-0.6843	0.08334	0.32819	0.07095	44.881
	(ZINB)	0.08773	-0.6847	0.08205	0.32806	0.07036	45.880
	(ZIG)	-0.2832	-0.59	0.1623	0.1099	0.2391	53.575
25	(ZIP)	-1.4194	-0.6718	0.3958	0.1969	0.3675	64.473
	(ZINB)	0.08773	-0.6847	0.08205	0.32806	0.07036	45.880
	(ZIG)	-1.3432	-0.7826	0.4339	0.367	0.2383	76.479
50	(ZIP)	-0.6449	0.05723	0.30813	-0.3019	0.30557	106.77
	(ZINB)	-0.6067	0.07636	0.30223	-0.3131	0.30306	107.62
	(ZIG)	-0.7065	0.02159	0.31286	-0.3655	0.37457	139.45
100	(ZIP)	-0.6492	-0.0695	0.3597	0.09444	0.01734	225.39
	(ZINB)	-0.6494	-0.0696	0.35977	0.09441	0.01739	226.40
	(ZIG)	-0.951	-0.2346	0.4235	0.1737	0.0558	292.26

Table 1 b: ESTIMATES OF PARAMETERS AND THE AIC FOR THE COUNT PART

sample (n)	MODEL						AIC
150	(ZIP)	-0.9222	-0.1738	0.40683	0.14862	0.0728	329.70
	(ZINB)	-0.9228	-0.1737	0.40695	0.14866	0.07281	330.69
	(ZIG)	-1.1435	-0.199	0.4508	0.1576	0.1087	422.50
300	(ZIP)	-0.8382	-0.1146	0.37551	0.11045	0.05904	1075.1
	(ZINB)	-0.8382	-0.1147	0.37554	0.11065	0.05894	1076.1
	(ZIG)	-1.0433	-0.1337	0.4187	0.13711	0.07907	1359.0
	(ZIP)	-0.8832	0.34639	0.40085	0.02528	-0.0371	1702.23

500	(ZINB)	-0.84	0.36778	0.39238	0.00553	-0.0364	1703.13
	(ZIG)	-1.0569	0.45453	0.44898	0.01929	-0.045	1924.56
1000	(ZIP)	0.15591	0.14495	0.02814	0.13906	0.11985	2123.2
	(ZINB)	0.15632	0.14471	0.02848	0.1395	0.11953	2124.2
	(ZIG)	0.24095	0.22059	0.05501	0.20946	0.17963	2706.6

In tables1a and 1b, we have the estimates of the parameters and the Akaike information criterion (AIC) for the count part at sample sizes of 15, 25, 50 and 100. The results shows that the parameters estimates of ZIP and ZINB has very close values when n=15, 50, 150, 300, 500 and when n= 1000. There was a slight disparity when n=25. By these results it showed the relationship between ZIP and ZINB. The estimates of the ZIG were not in any way similar to that of ZIP and ZINB because of the nature of ZIG model One of the important property of this distribution is the lack of memory property and in case of its truncation, situations arise practically in cases where the ability to record, or even to know about, occurrences is limited to values which lie above or below a given threshold or within a specified range.

Furthermore, the values AIC of the ZIP and ZINB were closely related unlike the AIC of ZIG which were far higher to that of ZIP and ZINB. However, the values of the AIC for ZIP were smaller than that of ZINB in all the sample sizes. These values indicated that ZIP outperformed he ZINB particularly when that zero-inflated parameter
$\omega = 0.1 \; and \; the \; vector \; of \; non-negative \; mean \; \lambda = 1.5$

Table 2a: ESTIMATES OF PARAMETERS FOR THE ZRO MODELS

sample (n)	MODEL	β_0	β_1	β_2	β_3	β_4
15	(ZIP)	47.066	-11.996	-16.453	5.336	-8.335
	(ZINB)	47.87	-6.76	-20.121	2.713	-8.028
	(ZIG)	47.794	-6.946	-18.982	2.995	-8.374
	(ZIP)	19.039	-15.877	-19.123	-10.909	9.573

25	(ZINB)	47.87	-6.76	-20.121	2.713	-8.028
	(ZIG)	24.58	-27.05	-38.1	-14.58	14.45
50	(ZIP)	72.96	-41.66	-63.18	-62.42	41.61
	(ZINB)	47.31	-11.85	-37.87	-32.48	16.85
	(ZIG)	53.4	-30.43	-46.77	-44.29	29.41
100	(ZIP)	33.184	28.234	-39.144	-32.979	5.442
	(ZINB)	31.06	26.151	-37.028	-30.907	5.449
	(ZIG)	22.769	5.56	-43.488	-22.536	7.985

Table 2b: ESTIMATES OF PARAMETERS FOR THE ZERO MODELS

sample (n)	MODEL	β_0	β_1	β_2	β_3	β_4
150	(ZIP)	25.9884	24.2166	-26.138	-17.416	-0.8887
	(ZINB)	28.0406	27.4256	-28.205	-19.529	-0.8942
	(ZIG)	36.426	41.259	-40.807	-36.743	3.545
300	(ZIP)	24.819	12.348	-26.597	-17.047	1.944
	(ZINB)	25.58	12.656	-27.351	-17.397	1.939
	(ZIG)	31.109	12.763	-36.081	-22.847	4.515
500	(ZIP)	7.229	-54.426	-65.474	2.646	16.791
	(ZINB)	24.062	-38.748	-48.638	1.22	8.933
	(ZIG)	11.0064	-54.116	-56.078	0.9726	17.0123
1000	(ZIP)	47.336	134.824	45002.9	54.4	47.223
	(ZINB)	13.517	17.883	23.221	5.402	6.153
	(ZIG)	15.6281	95.8137	125.528	32.9698	35.9045

Table 2a and 2b consist of the parameter estimates for the Zero part of the models./ Also closely related is the estimates between the ZIP, ZINB and ZIG when the sample size was 15.when sample size increases from 15 to 25 and other subsequent increase it showed that disparities exist among the estimates of the three models.

Table 3a: ESTIMATES OF STANDARD ERROR FOR THE COUNT PART

sample (n)	MODEL	β_0	β_1	β_2	β_3	β_4
15	(ZIP)	2.46021	0.6926	0.62872	0.79232	0.63549
	(ZINB)	4.53711	0.77268	1.14323	0.82297	0.69727
	(ZIG)	3.7662	1.1321	0.9512	1.2598	1.0297
25	(ZIP)	0.9228	0.587	0.1456	0.5665	0.5015
	(ZINB)	4.53711	0.77268	1.14323	0.82297	0.69727

	(ZIG)	1.4055	0.9429	0.2542	0.998	0.8534
50	(ZIP)	0.49113	0.40026	0.08207	0.36556	0.32889
	(ZINB)	0.45699	0.39434	0.08332	0.30861	0.2568
	(ZIG)	0.7524	0.65681	0.14932	0.61606	0.54811
100	(ZIP)	0.29445	0.23931	0.05479	0.23518	0.20329
	(ZINB)	0.2935	0.23927	0.05477	0.23384	0.20203
	(ZIG)	0.4649	0.4008	0.1069	0.355	0.3078

Figure 1.0: Plots of Standard Errors (Count Part) at n= 15, 25, 50 and 100

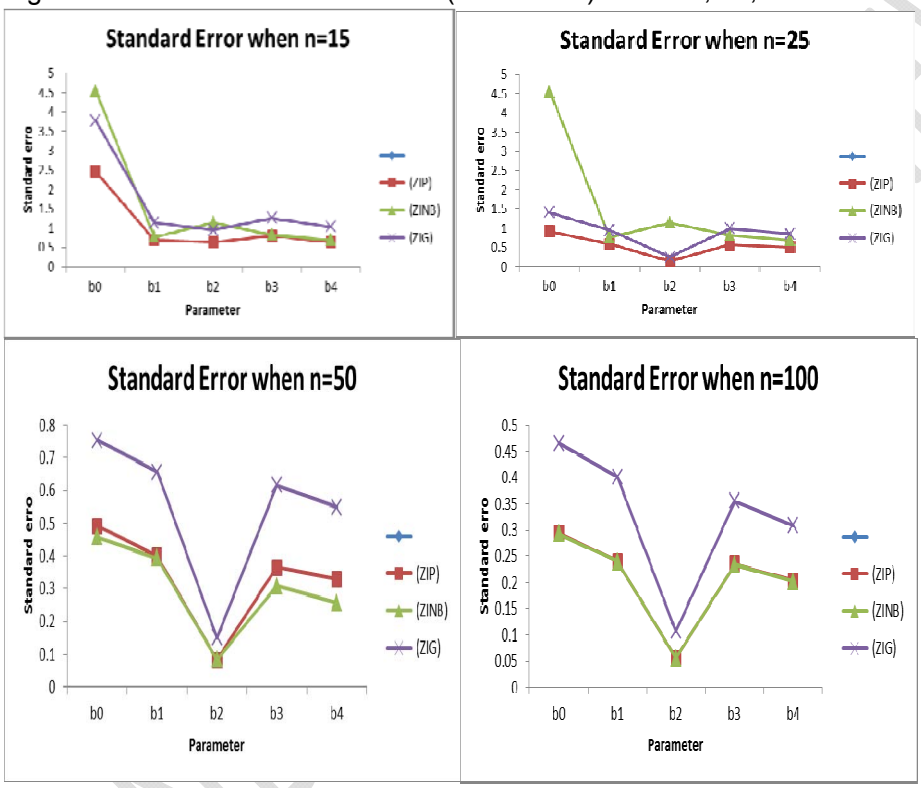
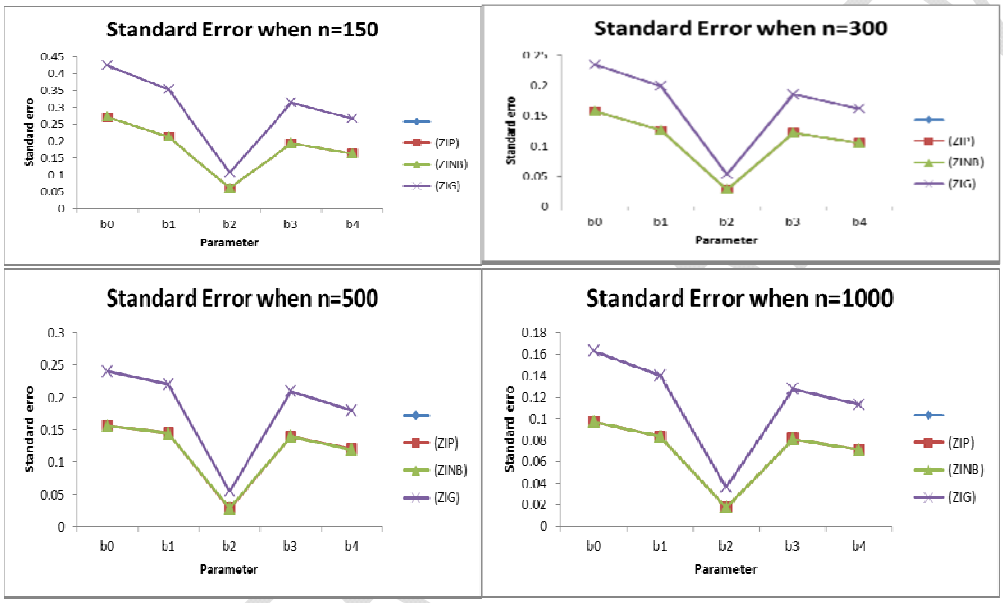


Table 3b: ESTIMATES OF STANDARD ERROR FOR THE COUNT PART

sample (n)	MODEL	β_0	β_1	β_2	β_3	β_4
150	(ZIP)	0.27174	0.21293	0.06137	0.19394	0.16428
	(ZINB)	0.27193	0.21294	0.06138	0.19407	0.16433
	(ZIG)	0.4245	0.3537	0.107	0.3146	0.2672
300	(ZIP)	0.15718	0.12618	0.02824	0.12228	0.10536
	(ZINB)	0.15728	0.12625	0.02826	0.1223	0.10536

	(ZIG)	0.23443	0.1993	0.05326	0.18545	0.16135
500	(ZIP)	0.15591	0.14495	0.02814	0.13906	0.11985
	(ZINB)	0.15632	0.14471	0.02848	0.1395	0.11953
	(ZIG)	0.24095	0.22059	0.05501	0.20946	0.17963
1000	(ZIP)	0.09743	0.08345	0.01766	0.08113	0.07145
	(ZINB)	0.09734	0.08345	0.01766	0.08113	0.07142
	(ZIG)	0.16288	0.14032	0.03649	0.12769	0.11344

Figure 2: Plots of Standard Errors (Count Part) at n= 150, 300, 500 and 1000



The standard errors (SE) of the models at the count part showed that ZIP has the least standard errors closely followed by ZINB. Though, at some points the SE of ZINB was higher than ZINB and ZIP almost equal ZIP when sample size was 25. From the plots, ZIP has the least SE as the plots fall below ZINB and ZIG at different sample sizes (see plots 1 and 2)

Table 4a: ESTIMATES OF STANDARD ERROR FOR THE ZERO PART

sample (n)	MODEL	β_0	β_1	β_2	β_3	β_4
15	(ZIP)	320.391	353.747	81.418	117.004	73.409
	(ZINB)	314.848	448.317	372.484	206.387	62.976
	(ZIG)	318.325	290.854	182.575	170.321	130.51
25	(ZIP)	185.753	164.901	187.018	197.766	152.048
	(ZINB)	314.848	448.317	372.484	206.387	62.976
	(ZIG)	697.01	2195.2	NA	1864.37	1809.89

50	(ZIP)	614.16	NA	272.88	599.52	NA
	(ZINB)	69.79	160.26	54.39	165.65	161.36
	(ZIG)	160.82	281.58	142.14	273.83	252.23
100	(ZIP)	107.555	98.436	116.115	107.714	19.79
	(ZINB)	70.382	56.211	82.974	71.336	19.873
	(ZIG)	78.243	3832.22	3648.11	3830.34	3826.95

Figure 3: Plots of Standard Errors (Zero Part) at n= 15, 25, 50 and 100

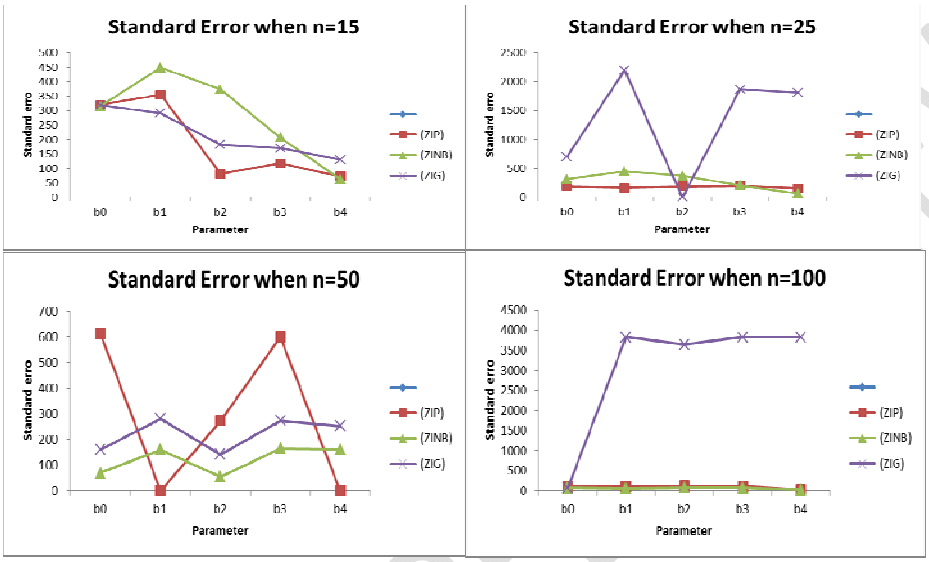
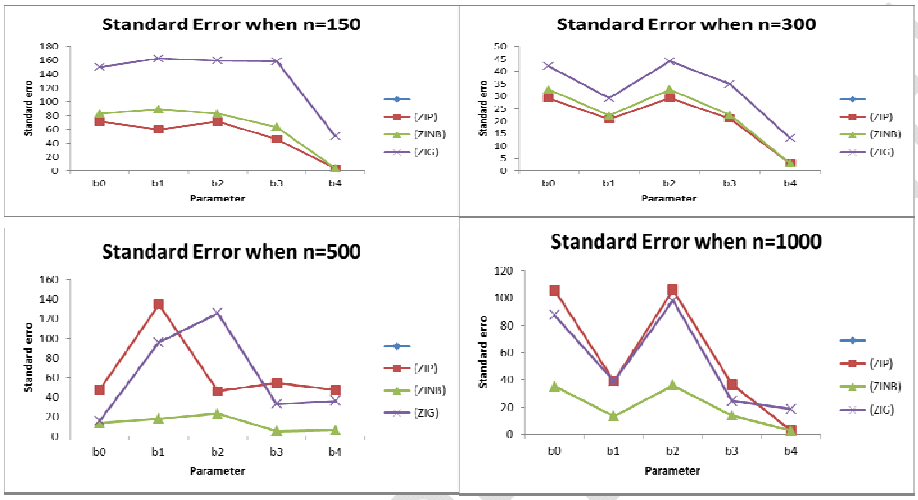


Table 4b: ESTIMATES OF STANDARD ERROR FOR THE ZERO PART

sample (n)	MODEL	β_0	β_1	β_2	β_3	β_4
150	(ZIP)	71.3294	60.0753	71.4479	46.0728	3.5575
	(ZINB)	82.7997	89.0968	82.899	63.4025	3.6493
	(ZIG)	150.145	162.593	159.746	158.588	49.976
300	(ZIP)	29.084	20.781	29.193	21.024	2.964
	(ZINB)	32.575	22.291	32.672	22.474	2.946
	(ZIG)	42.134	29.177	44.042	34.865	13.147

500	(ZIP)	47.336	134.824	45.9	54.4	47.223
	(ZINB)	13.517	17.883	23.221	5.402	6.153
	(ZIG)	15.6281	95.8137	125.528	32.9698	35.9045
1000	(ZIP)	105.686	39.1409	105.919	36.7716	3.1779
	(ZINB)	35.4013	13.4138	35.9497	13.9835	2.84322
	(ZIG)	87.479	38.871	98.01	24.435	18.619

Figure 4: Plots of Standard Errors (Zero Part) at n= 150, 300, 500 and 1000



The standard errors (SE) of the models at the zero part showed that ZIP has the least standard errors closely followed by ZINB. Though, at some points the SE of ZIP was higher than ZINB and ZIG. At sample sizes of 500 and 1000, ZINB outperformed the ZIP and ZIG (see plots 3 and 4).

CONCLUSION

Consequently, the plots of the standard errors and the AIC showed some disagreement on which model best fit at some points such as when sample size was 50, 100, 500 and 1000. The sizes of AICs for ZIP were lesser than ZINB which show superiority in terms of model fitting. However we can conclude that ZIP outperformed the ZINB in the Count part whereas ZINB at some points when the sample size rises above 300 outperformed the ZIP and ZIG in accounting for the zero part of the model.

Reference

1. Agresti, A. (1996). *An introduction to categorical data analysis*. New York: John Wiley

An application to domestic violence data. *Journal of Data Science, 4,* 117-130.

2. Bryk, A. S., Raudenbush, S. W., & Congdon, R. T. (1996). *HLM: Hierarchical linear* Cambridge University Press.Cameron, A. C., & Trivedi, P. K. (1998). Regression analysis of count data. New York: CB2 8RU, UK, *Published in the United States of America by Cambridge*

3. Colin. A and Trivedi (1998): Regression of Count Data: Cambridge University Press 0521632013
4. Colin. A and Trivedi (1999):Essentials of Count regression. Online Publication

5. Delucchi, K. L., & Bostrom, A. (2004). Methods for analysis of skewed data distributions in psychiatric clinical studies: Working with many zero values. *American Journal of Psychiatry, 161,* 1159-1168.
6. Famoye, F., & Singh, K. (2006). Zero-inflated generalized Poisson regression model with an Application to Domestic Violence Data. Journal of Data Science. 4(2006), 117-130
7. Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics, 34,* 1-14.
8. Zorn, C. J. W. (April 18-20, 1996). Evaluating zero-inflated and hurdle Poisson specifications. Midwest Political Science Association, 1-16.