International Journal of Engineering Inventions
e-ISSN: 2278-7461, p-ISSN: 2319-6491
Volume 15, Issue 11 [November 2026] PP: 58-65

Modelling Count Data with Over-Dispersion Using Double Poisson and Zero-Inflated Poisson Regressions Models with Application to Neo-Natal Mortality

[1]Mustapha Muhammad Lamido
[1,]Department of Mathematics and Statistics, Federal Polytechnic Bauchi, Nigeria

ABSTRACT
Neonatal mortality remains a major public health issue, particularly in developing countries, requiring robust statistical approaches to identify its determinants. This study compares Poisson Regression (PR), Zero-Inflated Poisson Regression (ZIP), and Double Poisson Regression (DPR) in modeling neonatal mortality data. The dependent variable measures neonatal deaths, while five independent variables capture patient diagnoses. Although PR assumes equidispersion, diagnostics revealed overdispersion and excess zeros, necessitating ZIP and DPR. ZIP accounts for zero inflation, while DPR flexibly handles both overdispersion and underdispersion. Model fit was evaluated using AIC and BIC, with DPR outperforming the other models by achieving the lowest values, indicating superior performance.

Keywords; Poisson Regression, Zero inflated poisson, Double poisson, Neonatal Mortality

---

Date of Submission: 02-11-2026 Date of acceptance: 12-11-2026

I. INTRODUCTION

The Poisson distribution serves as the benchmark model for analyzing count data, which comprises observations restricted to non-negative integers 0, 1, 2, ... Count data are inherently discrete, often skewed, typically associated with rare events, and may exhibit characteristics such as overdispersion and excess zeros. Unlike ordinal data, these integers arise from counting processes rather than ranking. An event count represents the realization of a nonnegative integer-valued random variable and reflects the frequency of an occurrence for example, the number of traffic accidents, earthquakes, or dengue fever cases by Winkelmann. R, (2008). Various statistical models have been developed to address the unique features of count data, with the appropriate choice depending on the underlying distributional properties. In regression settings, count outcomes are generally modeled as dependent variables to examine their relationship with explanatory factors.

Before selecting an appropriate modeling technique, it is essential to first understand the nature of the data. When dealing with discrete, non-negative integer outcomes, applying least squares regression is often unsuitable, as it presents methodological and statistical limitations according to Mouatassim Y. and Ezzahid, E. (2012). Unlike classical linear regression, count regression models are inherently nonlinear and are designed to handle characteristics specific to count data, such as discreteness, nonlinearity, and restriction to non-negative values. Poisson regression is widely regarded as the starting point for modeling count data, as it aligns well with these properties. Common examples of count outcomes include the number of doctor visits by patients, patents awarded to firms, road accident fatalities, or dengue fever cases, variables typically recorded as small, non-negative integers by Camaron, A.C and Trivedi P. K(2008). For such data, Poisson regression offers clear advantages over conventional linear regression models, Wedderbum, R.W.M.(1974). Nevertheless, the Poisson model has an important limitation: it assumes equality between the mean and variance, a condition known as equidispersion. When this assumption does not hold for instance, when the variance of observed counts exceeds the mean overdispersion arises. Ignoring overdispersion may result in inconsistent estimates, underestimated standard errors, and inflated test statistics.

In count data analysis, it is common practice to begin with the Poisson regression model and then adjust for overdispersion if detected. One method for addressing overdispersion is the Quasi-Likelihood estimation technique proposed by Camaron, A.C and Trivedi.P. K (2005). Another widely used approach is the Negative Binomial (NB) regression model,which generalizes the Poisson-gamma regression framework by Hilbe J.M (2011). The NB model introduces a dispersion parameter that allows the variance to exceed the mean, thereby accounting for both overdispersion and unobserved heterogeneity in the data. Beyond overdispersion, real-world count data often display excess zeros. While the NB model effectively handles overdispersion, specialized models such as the Hurdle model and Zero-Inflated Poisson (ZIP) or Zero-Inflated Negative Binomial (ZINB) models are better suited for managing excessive zero counts.

II. MATERIALS AND METHOD

Let Y_i be a random variable that takes non-negative integer values, i = 0, 1, 2,…, n. We assume that this random variable represents count observations that follow a probability distribution consistent with the framework of a Generalized Linear Model (GLM). The relationship is expressed through a multiple linear regression model with an appropriate link function, formulated as:

$$\mu_i = e^{X_i^T \beta_i} \tag{1}$$

Therefore,

$$\mu_i = [1\ X_1\ X_2\ X_3\ X_4\ X_5]^T \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{bmatrix}, \tag{2}$$

This implies,

$$\mu_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 \tag{3}$$

where

- $X_i^T \beta_i$ is a Multiple linear regression model.
- X_i^T is a k-dimentional column vector of covariate.
- β_i, $i = 1,2,3,4,5$ represent effect of the covariates $X_1, X_2 \ldots X_5$ on the response variable Y
- i represent i^{th} observation usually included in the set of regressors X_i^T for the intercept.
- β_0 is the intercept term usually added in the set of parameters

Poisson Regression (PR)

Assuming that the independent random variable Y_i is count data that follows Poisson Regression model written as

$$Pr(Y = y_i) = \frac{e^{\lambda_i} \lambda_i^{y_i}}{y_i!}, For \lambda_i > 0 y_i = 0,1,2,\ldots \tag{4}$$

With mean and the variance equal

$$E(Y) = \lambda_i = Var(Y) \tag{5}$$

The poison linear regression model is formulated through the mean, such that

$$\ln(\lambda) = X_i^T \beta, \tag{6}$$

Where $X_i^T \beta_i$ is given in equation (6). The regression parameter β_i is estimated by the maximum likelihood.

Maximum Likelihood Estimation for Parameters of Poisson

Suppose that $Y_1, \ldots, Y_n$ are independent Poisson random variables with common density where β is an unknown location parameter describing the centre of the distribution. The likelihood equation for MLE is given by

$$\ln L(\beta) = \sum_{i=1}^{n} \left(-\lambda_i + y_i In\lambda_i - In(y_i!) \right) \tag{7}$$

Given the Link function $\lambda_i = e^{x_i \beta}$, we have

$$\ln L(\beta) = \sum_{i=1}^{n} \left(\left(y_i \left(X_i^T \beta \right) - e^{\left(X_i^T \beta \right)} - In(y_i!) \right) \right) \tag{8}$$

by differentiating equation (10) with respect to β, we have

$$\frac{\partial \ln L}{\partial \beta_k} = \sum_{i=1}^{n} \left(\left(y_i - e^{(x_i^T \beta)} \right) X_i \right) = 0 \tag{9}$$

the Hessian of this function is:

$$\frac{\partial^2 \ln L}{\partial\beta\partial\beta'} = \sum_{i=1}^{n}\left(X_i^T X_i e^{X_i^T\beta} \right) \tag{10}$$

The Newton-Raphson method can be applied to generate a sequence that converges to the MLE.If we assume β as a $k*1$ vector, we can iterate

$$\beta_{(i+1)} = \beta_{(i)} \quad \frac{\frac{\partial \ln L}{\partial\beta_k}}{\frac{\partial^2 \ln L}{\partial\beta\partial\beta_i'}} \tag{11}$$

where $\frac{\partial \ln L}{\partial\beta_k}$ is the gradient vector of the log-likelihood function, and $\frac{\partial^2 \ln L}{\partial\beta\partial\beta_i'}$ is the Hessian of the log-likelihood function.

Zero Inflated Poissson Regression (ZIPR)

The model introduced by Lambert, D. (1992). was designed to handle count data characterized by excess zeros. In Zero-Inflated Poisson (ZIP) regression, The zero counts is assumed to occur with p_i as probability , while with probability$1-p_i$, the counts follow a Poisson distribution with mean λ_i , where i = 0,1,2 ,…n Accordingly, the ZIP model can be viewed as a mixture of two components: one representing the structural zeros and the other representing the Poisson-distributed counts, expressed as:

$$P(Y = y_i \vee \lambda, p_i) = \begin{cases} p_i + (1+p_i)e^{\lambda_i}, y_i = 0 \\ (1 \quad p_i)\frac{e^{\lambda_i}\lambda_i^{y_i}}{y_i!}, y_i = 1,2,... \end{cases} \tag{12}$$

The first term of the equation represents the zero-inflation component, while the second term corresponds to the non-zero count component. Together, these two parts define the Zero-Inflated Poisson (ZIP) model. The mean and variance of the ZIP distribution are given as follows:

$$E(Y_i) = \lambda_i(1 \quad p_i) \tag{12}$$

and

$$Var(Y) = \lambda_i(1 \quad p_i)(1+\lambda_i p_i)^2 \tag{13}$$

It is important to note that zero outcomes in the ZIP framework can originate from both the zero-inflation component and the Poisson component. The zero-inflation part specifically accounts for the 'excess' zeros observed beyond those expected under a standard Poisson distribution. To evaluate the influence of covariates on the count process within the ZIP model, the model, p_i *and* λ_i are expressed as functions of covariates through suitable link functions. A common approach for modeling the probability of excess zeros is logistic regression with a logit link, formulated as:

$$logit(p_i) = z_i\delta \tag{14}$$

$$p_i = \frac{e^{z_i\delta}}{1+e^{z_i\delta}} \tag{15}$$

$$p_i = \frac{1}{1+e^{z_i\delta}} \tag{16}$$

The effect of covariates count data without excess zeros can be modelled through Poisson regression:

$$\ln(\lambda) = X_i^T\beta_i, \tag{17}$$

where $X_i^T\beta_i$ is given in equation (6). The regression parameter β_i are estimated by the maximum likelihood.

Double Poisson Regression (DPR)

Came from the double exponential family, Efron, B. (1986) introduced the Double Poisson distribution. This model is characterized by two parameters, μ and θ. Zou, Y., et al (2013) provided an approximate probability mass function (p.m.f.) expressed as:

$$P(Y=y) = f_{\mu,\theta}(y) = c(\mu,\theta)\left(\theta^{\frac{1}{2}}e^{-\theta\mu}\right)\left(\frac{(e^{-y})(y^{y})}{y!}\right)\left(\frac{ey}{y}\right)^{\theta y}\ y = 0,1,2,\ldots, \quad (18)$$

where θ is the dispersion parameter, $\mu = exp(X_i\beta)$, β is a vector of regression parameters, and X_i is the covariate matrix. The exact double Poisson density is given as:

$$P(Y=y) = f_{\mu,\theta}(y)c(\mu,\theta)f_{\mu\theta}(y) \quad (19)$$

where the factor $c(\mu,\theta)$ to be calculated as:

$$c(\mu,\theta) = \sum_{y=0}^{\infty} f(\mu,\theta) \approx 1 + \frac{1}{12\mu\theta}\left(1+\frac{1}{\mu\theta}\right) \quad (20)$$

this denotes $c(\mu,\theta)$ the normalizing constant, which is approximately equal to 1. This constant $c(\mu,\theta)$ guarantees that the probability density sum to unity. The expected value and variance, with respect to the exact density $f_{\mu,\theta}$, are given as:

$$E(Y) \approx \mu \quad (21)$$

$$Var(Y) \approx \frac{\mu}{\theta} \quad (22)$$

The Double Poisson model accommodates both overdispersion (θ > 0) and underdispersion (θ < 0). When (θ = 1), the distribution reduces to the standard Poisson model. A key feature of the Double Poisson (DP) model is the inclusion of a normalizing constant that involves an infinite series by Zou, Y., et al (2013) expressed as:

$$\sum_{y=0}^{\infty} f_{\mu\theta}(y) \quad (23)$$

The Double Poisson Regression DPR model can be formulated through the mean:

$$E(Y_i) = \mu = exp(X_i^T\beta) \quad (24)$$

where $X_i^T\beta_i$ is given in equation (6). The regression parameter β_i and θ are estimated by the used of maximum likelihood.

Goodness of fit test

The adequacy of the Poisson regression model can be assessed by first examining basic descriptive measures of the count data. In the Poisson framework, the mean and variance are expected to be equal; a large discrepancy between them indicates possible overdispersion or, less commonly, underdispersion. Model diagnostics often use a scale parameter, calculated as the Pearson chi-square divided by the degrees of freedom (observations minus parameters), to measure dispersion. In practice, underdispersion is seldom observed in count data.

Akaike Information Criterion (AIC)

Akaike Information Criterion (AIC) provides a way of selecting a model from a set of models. It is defined according to Akaike H. (1973). as:

$$AIC = 2k \quad 2\log(L) \quad (24)$$

where:

- k is the number of estimated parameters for the model.
- L is the maximum value of the likelihood function for the model.

For a given dataset, the model with the smallest AIC value is chosen as the most appropriate compared to the alternatives.

Bayesian Information Criterion (BIC)

The Bayesian Information Criterion(BIC) introduced by Schwarz, and Gideon E. (1978), is a model selection technique used to identifythe most suitable model for a dataset from other models. The BIC is mathematically computed using:

$$BIC = k\log(n) \quad 2\log(L) \quad (25)$$

where:

- k is the number of parameters estimated for the model.
- n is the number of observations.
- L is the maximized value of the likelihood function for the model.

The model with the lowest BIC value is consider as the appropriate model for a given data set among other models.

III. RESULT AND DISCUSSION

The Numerical results of the models examined are displayed in terms of Tables, plots and graphs accordingly with the response variable as the number of Neonatal mortality (NNM) and that of categorical variables are Preterm Low Birth Weight (PLBW), Severe Birth Asphyxia (SBA), Small Gestational Age (SGA), Neonatal Jaundice (NNJ) and Obstetric Complication respectively. Before starting the model formulation, it will first be tested whether the response variable in this case, that is the number of neonatal mortality is following a Poisson distribution or not.

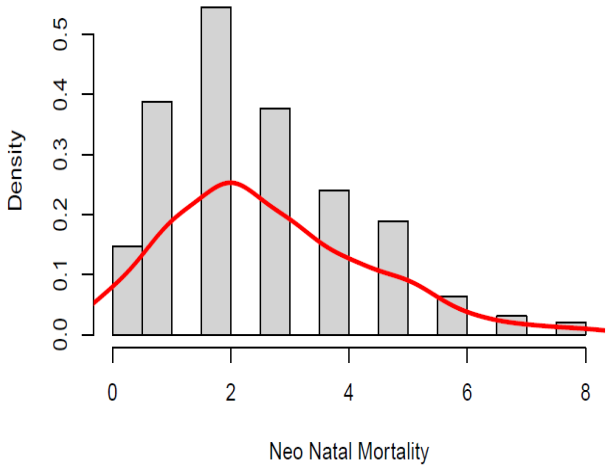


Figure 3.1: Histogram of the Neo natal mortality

From figure 3.1 above, we can see that the histogram of the Neo natal mortality data is not in bell shape meaning that the data did not follow a normal distribution. Therefore, we can go ahead and model the count data using the exponential family distributions.

Table 3.1 Descriptive statistics of Neo-natal mortality count data from 2018-2021

Variable	Mean	Variance
NNM	2.638743	2.884596

Source: Computed by the authors using R software version 4.1.1.

Table 3.1 the descriptive statistics for the mean and variance of neonatal mortality count data and its determinant factors from 2018–2021 are presented. The results indicate that the variance exceeds the mean, suggesting the presence of overdispersion in the dataset.

Table 3.2 Analysis of Neo natal mortality using Poisson regression(PR)

	Poisson regression(PR)			
Parameter	Estimates β	Std. Error	Z-Value	P-Value
Intercept	-1.3651	0.1590	-8.683	< 2e-16 ***
PLBW	0.8418	0.1543	5.545	2.92e-08 ***
SBA	1.0354	0.1668	6.209	4.33e-10 ***
AGA	0.7316	0.1706	4.289	1.79e-05 ***
NNJ	0.9471	0.1742	5.436	4.45e-08 ***
OC	0.9907	0.1456	6.806	1.00e-11 ***

Source: Computed by the authors using R software version 4.1.1.Note: (***) denote significance at 1% respectively.

The Poisson Regression model:

$$\mu = -1.3651 + 0.8418X_1 + 1.0354X_2 + 0.7316X_3 + 0.9471X_4 + 0.9907X_5$$

Table 3.3 Result of Variance inflation factor(VIF) of Poisson model

PLBW	SBA	AGE	NNJ	OC
1.032137	1.223795	1.196812	1.061956	1.072765

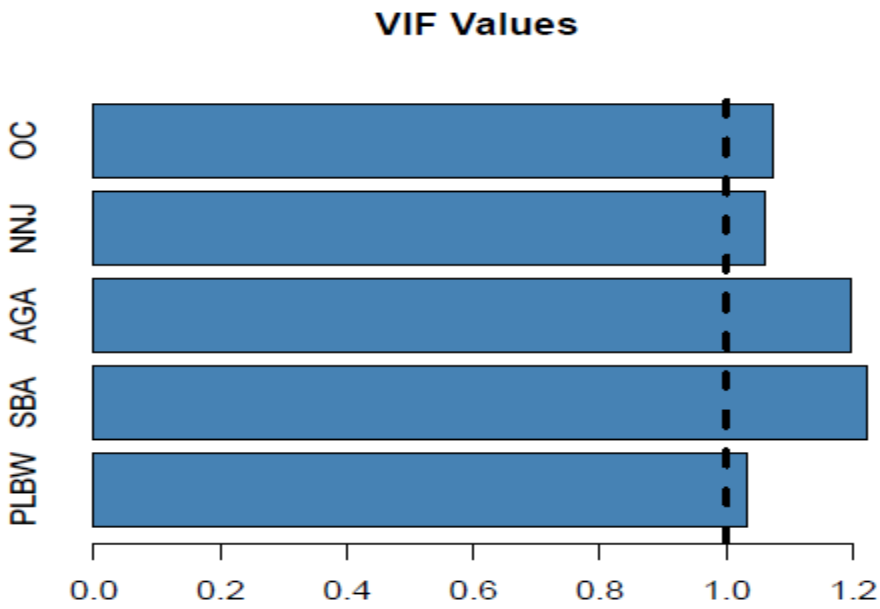


Figure 3.2 Graph of VIF Values of the Poisson model

Table 3.4 Analysis of Neo natal mortality using Zero-Inflated Poisson regression(ZIPR)

Count model				
Parameter	Estimates β	Std. Error	Z-Value	P-Value
Intercept	-0.5113	0.1986	-2.575	0.010024 *
PLBW	0.5004	0.1598	3.131	0.00174 **
SBA	0.6391	0.1685	3.793	0,000149 ***
AGA	0.4861	0.1684	2.886	0.003900 **
NNJ	0.5227	0.1832	2.853	0.004332 **
OC	0.5536	0.1599	3.462	0.000537 ***
Zero-inflated model				
Parameter	Estimates δ	Std. Error	Z-Value	P-Value
Intercept	-15.15	332.61	0.046	< 0.964
PLBW	-28.66	429.85	-0.067	0.947
SBA	-41.58	173474.09	0.000	1.000
AGA	-29.47	610.41	-0.048	0.961
NNJ	-28.71	440.86	-0.065	0.948
OC	-28.67	390.20	-0.072	0.942

Source: Computed by the authors using R software version 4.1.1.Note: (***) denote significance at 1% respectively.

The count model of ZIPR:

$$\mu = -0.5113 + 0.5004X_1 + 0.6391X_2 + 0.4861X_3 + 0.5227X_4 + 0.5536X_5$$

Table 3.5 Analysis of Neo natal mortality using Double Poisson regression(DPR)

Double Poisson regression (DPR)				
Parameter	Estimates β	Std. Error	Z-Value	P-Value
Intercept	-1.20019	0.0535	-22.456	< 2e-16 ***
PLBW	0.76769	0.06097	12.591	< 2e-16 ***
SBA	0.96925	0.06635	14.609	< 2e-16 ***
AGA	0.67850	0.06859	9.892	< 2e-16 ***

Double Poisson regression (DPR)				
NNJ	0.81088	0.07032	11.531	< 2e-16 ***
OC	0.86356	0.05748	15.038	< 2e-16 ***
	Dispersion θ			
	0.97889207			

Source: Computed by the authors using R software version 4.1.1.Note: (***) denote significance at 1% respectively.

The Double Poisson Model:

$$\mu = -1.20019 + 0.76769X_1 + 0.96925X_2 + 0.67850X_3 + 0.81088X_4 + 0.86356X_5$$

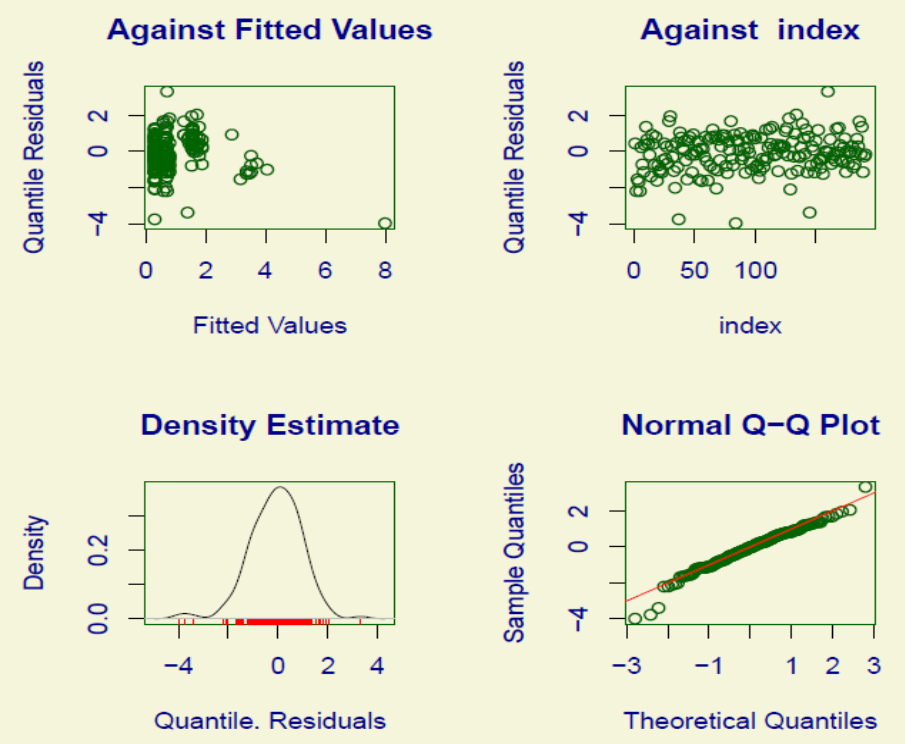


Figure 3.4: Diagnostic plot of DPR Residual

Table 3.6 Analysis of goodness of fit for the models examined

Models	Df	AIC	BIC
PR	6	340.5506	360.0643
ZIP	12	296.5282	335.5555
DPR	7	177.4135	200.1794

Source: Computed by the authors using R software version 4.1.1.

IV. DISCUSSION

Table 3.2 shows the Poisson regression (PR) analysis of neonatal mortality in Gombe State. Table 3.3 gives the multicollinearity test results for the Poisson model. Since the maximum likelihood estimator (MLE) can become unstable if multicollinearity exists, it is important to check for milticollineariaty after fitting the model. The VIF test in Table 3.3 shows that all predictor variables have VIF values below 2, meaning there is no serious multicollinearity. Table 3.4 then shows the results of neonatal mortality analysis using the Zero-Inflated Poisson (ZIP) model, covering both the count and zero-inflated parts. Furthermore, Table 3.5 presents the analysis of neonatal mortality rates using the Double Poisson Regression (DPR) model, which serves as an alternative approach for handling the problem of over-dispersion in the data. Finally, Table 3.6 reports the goodness-of-fit assessment for the models considered, based on a comparison of their AIC and BIC values. The model with the lowest AIC and BIC is identified as the best-fitting model for the neonatal mortality data. From the AIC and BIC values of the three models, it is evident that the Double Poisson Regression (DPR) provides the best fit, with the smallest AIC (177.4135) and BIC (200.1794), followed by the Zero-Inflated Poisson Regression (ZIPR) with AIC (296.5282) and BIC (335.5555), and lastly the Poisson Regression (PR) with AIC (340.5506) and BIC (360.0643). The consistently lower AIC and BIC values of the DPR model indicate its superior performance compared to the other models. This suggests that the DPR is the most appropriate model

for the data, as it can effectively handle both over-dispersion and under-dispersion. In this case, the DPR adequately addressed the over-dispersion present in the neonatal mortality data, which in turn enhanced the significance of the predictor variables. Furthermore, the diagnostic plots in Figure 3.4 confirm how well the DPR model fits the neonatal mortality data.

V. CONCLUSIONS

From the findings of this study, it can be conclude that the Zero-inflated Poisson regression cannot handle over dispersion problem, as it was seen in the neo natal mortality analysis. Likewise, we have seen that due to present of zeros in the data set. The zero-inflated Poisson regression model can outperform the standard Poisson regression. The study, also found that Zero inflation model cannot handle some effect of over-dispersion.Therefore, this study concluded that the appropriate exponential distribution method to be used for

modelling the Neo natal mortality rate in Nigeria and similar data, is to adopt the used of double poison regression model. Because, the DPR model can handle both over/under dispersion property and when there is equidispersion in the data, it will automatically collapse to standard Poisson regression.

REFERENCES

[1]. Cameron, A.C., and Trivedi P.K. (1998). Regression Analysis of Count Data, New York:Cambridge University Press.
[2]. Winkelmann, R. (2008). Econometric Analysis of Count Data., 5th edition. Springer-Verlag,Berlin.
[3]. Akaike H. (1973). Information theory and an extension of the maximum likelihood principle. In: Petrov, B. N. and Csàki, F. (eds): Second international symposium on inference theory, Budapest, Akadèmiai Kiadó, pp. 267-281.
[4]. Consul, P. C. and Famoye, F. (1992). Generalized Poisson regression model. Comm. Statist.Theory Methods 21, 89-109.
[5]. Mouatassim Y. and Ezzahid E., (2012). Poisson regression and Zero-inflated Poisson regression: application to private health insurance data. Eur. Actuar. J. DOI 10.1007/s13385-012-0056-2.
[6]. Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. Biometrika 61:439–447.
[7]. Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing technometrics, 34, 1-14.
[8]. Hilbe J. M. (2011). Negative binomial regression. 2nd edition. Cambridge: Cambridge University Press.
[9]. Zao, Y., Geedipally, S. R., and Lord, D. (2013). Evaluating the Double Poisson Generalized Linear Model. Accident Analysis and Prevention, 59, 497-505.
[10]. Schwarz, and Gideon E. (1978), *"Estimating the dimension of a model", Annals of Statistics, **6** (2): 461–464, doi:10.1214/aos/1176344136, MR 0468014*.
[11]. Efron, B. (1986). Double Exponential Families and Their Use in Generalized Linear Regression.Journal of American Statistical Association, 81(295), 709-721.

ACKNOWLEDGEMENTS

I will like to thank the management of the Federal Polytechnic, sBauchi, the Research Directorate and the Tertiary Education Trust fund (TETFUND) for their support in making this study a success.