International Journal of Mathematical Archive-9(6), 2018, 201-212
Available online through www.ijma.info ISSN 2229 – 5046

A WALD TEST FOR OVER DISPERSION IN ZERO-INFLATED POISSON REGRESSION MODEL

CH. SREELATHA*[1] AND Dr. B. MUNISWAMY[2]

[1,2]Department of statistics, Andhra University, Visakhapatnam-530003, India.

(Received On: 20-04-18; Revised & Accepted On: 26-05-18)

ABSTRACT

W*e have considered the regression models to fit the count data especially in the field of Biometrical, Environmental, Social Sciences, and Economical areas. In the field of medical applications the count data with extra zeros are also common. The zero-inflated Poisson (ZIP) regression model is helpful to examine such data. Here we awareness on the use of ZIP model for analysis of count data including maximum likelihood estimation for regression coefficients using Fisher scoring method, compare between Poisson and ZIP models by various tests: likelihood ratio test, score test, chi-square test, test based on a confidence interval test and Cochran test. Model selection using Deviance method, AIC and BIC. We can find a Wald test for ZIP model in a single sample case for detecting zero-inflation in Poisson model and conduct a small simulation study in order to investigate sampling distribution of Wald test and power of Wald test. From our study we found that distribution can be used to detect the zero-inflation in counts.*

Keywords: *Count data, Poisson model, Zero-inflated Poisson, Fisher Scoring Method, Wald test, AIC, BIC.*

1. INTRODUCTION

In many medical applications, counts of events occurring in a given time or exposure period are discrete random variable having Poisson distribution. Based on the basic concept of generalized linear models (*glms*), the relationship between explanatory variables and response variable of Poisson counts can be described by Poisson regression or log linear models. The Poisson model is formed under two principal assumptions: one is that events occur independently over given time or exposure period and the other is that the conditional mean and variance are equal. However, in practice, the equality of the mean and variance rarely occurs; the variance may be either greater or less than the mean. If the variance is greater than the mean, it means that counts are more variable than specified by the Poisson events and are describe as overdispersion. If the variance is less than the mean, it means that counts are less variable than specified by the Poisson events and are describe as underdispersion. However, in practice, underdispersion is less common (McCullagh and Nelder, 1989). Moreover, overdispersion can be caused by excess number of observed zero counts, since the excess zeros will give smaller conditional mean than the true value. The count data with excess zeros, is known as zero-inflated Poisson counts. Of course it is possible to have fewer zero counts than expected, but this is less common in practice (Ritout *et al.*, 1998). The count data encountered contain excess zeros relative to the Poisson distribution. A popular approach to analyse such data is to use a zero-inflated Poisson (ZIP) regression model. The ZIP model combines the Poisson distribution with a degenerate component of point mass at zero. A review of the ZIP methodology can be found in References Often the outcome of interest in public health and medical studies is a count variable, which is usually assumed to follow the Poisson distribution and can be modeled accordingly. However, the count variable may contain excess zeroes above what is to be expected from the Poisson model. These excess zeroes may be due to 1) the presence of a subpopulation with only zero counts, 2) overdispersion, or 3) chance (Campbell, Machin, and D'Arcangues, 1991). A common approach for analyzing such data is the zero-inflated Poisson (ZIP) model (Mullahy, 1986; Lambert, 1992), which is an extension of Cohen's (1960) modified Poisson distribution. ZIP modeling has been used to analyze data on caries prevention (Bohning, Dietz, Schlattmann, Mendonca, and Kirchner, 1999), early growth failure in children (Cheung, 2002), and sudden infant death syndrome (Dalrymple, Hudson, and Ford, 2003), as well as many other diseases. ZIP modeling also can accommodate the extent of individual exposure (Lee, Wang, and Yau, 2001). ZIP models also have been extended to the bivariate (Walhin, 2001) and multivariate settings (Li, Lu, Park, Brinkley, and Peterson, 1999). Hall (2000) and Yau and Lee (2001) have used random-effects approaches to extend the ZIP model to analyze longitudinal count data. This paper will be focussed only on ZIP models and will proposed a Wald test for comparing between the standard Poisson and ZIP models.

Corresponding Author: Ch. Sreelatha*[1],
[1]Department of statistics, Andhra University, Visakhapatnam-530003, India.

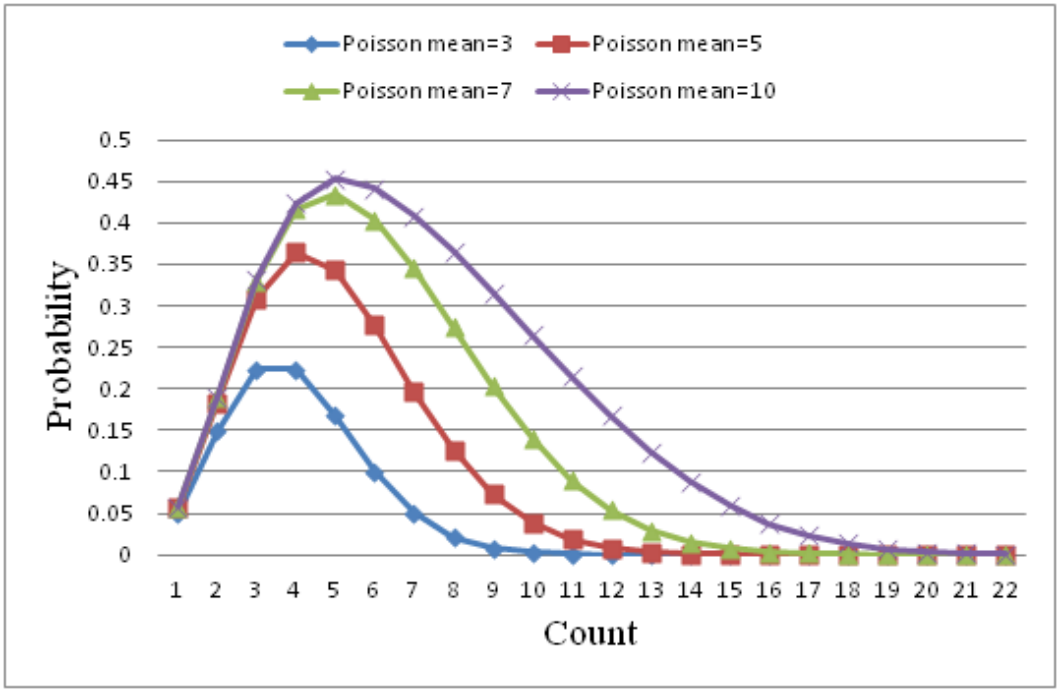
2. POISSON REGRESSION MODEL

Poisson regression model provide a standard framework for the analysis of count data. Suppose we have an independent sample on n pairs of observations $(y_i, x_i)_{i\in 1,2,....n}$, where y_i denotes the number of times an event has occurred and x_i is the value of the explanatory variable for the i^{th} subject. Assume $y_i \sim Poisson(\lambda_i)$ then the probability density function of Poisson random variables, y_i, is given by

$$f(y;\lambda) = \frac{e^{-\lambda}\lambda^{y}}{y!}, \quad y = 0,1,2,....$$

Where λ>0, represents the expected number of occurrences in a fixed period of time. The mean and variance of the Poisson regression model is give as follows:

Figure2.1 : Poisson distribution plot with different means



Figures-2.1: illustrated four characteristics of the Poisson distribution that are important to understand the regression models for count dataset. Then we have the following empirical observations.

1. As λ increases, the curve of the distribution shifts to the right, where is the mean of the distribution.
2. Var(y) = λ, which is known as equi-dispersion. In real data, many count variables have a variance greater than the mean, which is called Overdispersion.
3. As λ increases, the probability of a zero count decreases. For many count dataset, there are more observed zeros than predicted by the Poisson distribution.
4. As λ increases, the Poisson distribution approximates a normal distribution. This is shown by the distribution for λ=10.

Let x be $n\times p$ matrix of explanatory variables. The relationship between y_i and i^{th} row vector of x, x_i, linked by $g(\lambda_i)$ is given by:

$$\log(\lambda_i) = x_i^T\beta, i = 1,2,..., p$$

This model is known as the Poisson regression model or log-linear model. To find the maximum likelihood function of λ_i, we define the likelihood function as fallows:

$$L = l(\lambda_i, \beta) = \prod_{i=1}^{n} \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!} = \prod_{i=1}^{n} \frac{e^{-e^{x_i^T\beta}}(e^{x_i^T\beta})^{y_i}}{y_i!}$$

Taking log on both sides and we get:

$$L = \log(l(\lambda_i, \beta)) = \sum_{i=1}^{n}[y_i \log\lambda_i - \lambda_i - \log y_i!] = \sum_{i=1}^{n}\left[y_i x_i^T\beta - e^{x_i^T\beta} - \log y_i!\right]$$

The first derivative of the log likelihood function is

$$\frac{\partial L}{\partial \beta_j} = \sum_{i=1}^{n} \left[y_i - \exp(x_i^T \beta) \right] x_{ij}$$

The second derivative of the log likelihood function is

$$\frac{\partial^2 L}{\partial \beta_j \partial \beta_k} = -\sum \exp(x_i^T \beta) x_{ij} x_{ik}$$

Hence, the log-likelihood function of the Poisson regression model is nonlinear in β so that they can be obtained via using an iterative algorithm. The most commonly used iterative algorithms are either Newton-Raphson or Fisher scoring. In practice $\widehat{\beta}$ is the solution of the estimating equations obtained by differentiating the log likelihood, (2) in terms of β and equating them to zero. Therefore, β will be obtained by maximizing using numerical iterative method (McCullagh and Nelder, 1989).

3. ZERO-INFLATED POISSON REGRESSION MODEL

The zero-inflated Poisson model is a simple mixture model for count data with excess zeros, discovered by Lambert (1992). The model is a combination of a Poisson distribution and a degenerate distribution at zero. Specifically, if Y_i are independent random variables having a Zero-inflated Poisson distribution, the zeros are assumed to arise in two ways corresponding to distinct underlying states. The first state occurs with probability π_i and produces only zeros, while the other state occurs with probability 1- π_i and leads to a standard Poisson count with mean λ_i and hence a chance of further zeros. In general, the zeros from the first state are called structural zeros and those from the Poisson distribution are called sampling zeros. This two-state process gives a simple two-component mixture distribution with probability mass function.

$$P(y_i \mid \pi_i, \mu_i) = \begin{cases} \pi_i + (1-\pi_i)\exp(-\mu_i), & y_i = 0 \\ (1-\pi_i)\dfrac{\exp(-\mu_i)\mu_i^{yi}}{y_i!}, & y_i = 1,2,,..., 0 \le \pi_i \le 1 \end{cases} \quad (3.1)$$

Which we denote by $Y_i \sim ZIP(\mu_i, \pi_i)$. The mean and variance of Y_i are

$$E_{zip}(y_i \mid \pi_i, \mu_i) = (1-\pi_i)\mu_i \ ,$$

and

$$Var_{zip}(y_i \mid \pi_i, \mu_i) = E_{zip}(y_i \mid \pi_i, \mu_i)(1+\pi_i\mu_i)$$

For a random sample of observations $y_1, y_2, \ldots, y_n$, the log-likelihood function is given by

$$\ell = \ell(\mu, \pi; y) = \sum_{i=1}^{n} \left\{ \ln[\pi_i + (1-\pi_i)e^{-\mu_i}] I_{(y_i=0)} + [\ln(1-\pi_i) - \mu_i + y_i \ln \mu_i - \ln(y_i!)] I_{(y_i>0)} \right\} \quad (3.2)$$

Where I(.) is the indicator function for the specified event, i.e. equal to 1 if the event is true and 0 otherwise. To apply the zero-inflated Poisson model in practical modeling situations, Lambert(1992) suggested the fallowing joint models for μ_i and π_i

$$\ln(\mu_i) = x_i^T \beta \ and \ \ln\left(\frac{\pi_i}{1-\pi_i}\right) = z_i^T \gamma, \quad i = 1,2,...,n \quad (3.3)$$

Where, x_i and z_i are covariate matrices and β, γ are (p+1)×1 and (q+1)×1 vector of unknown parameters, respectively. Note that the vector of covariates x_i and z_i can be the same or different.

Clearly, the ZIP distribution, reduces to the Poisson distribution, when $\pi_i = 0$. However, $\pi_i = 0$ is not valid in the logit link function. Other link functions can be specified for either μ_i and π_i . while the logit link for π_i may seem the natural choice, fallowing the description given in Jansakul and Hinde (2002), it is useful to use the identity link giving the joint models

$$\ln(\mu_i) = x_i^T \beta \ \ and \ \ \pi_i = z_i^T \gamma. \quad (3.4)$$

Using an identity link allows the zero-inflated model to be extended to include possible zero-deflation, although the model fitting may need to be constrained.

3.1: Maximum likelihood estimation for ZIP regression models

Based on the ZIP model (3.1), the log-likelihood function(3.2) and the model for μ and π displayed in (3.4), it is obvious that being a finite mixture the ZIP distribution is not a member of exponential family distribution and so standard GLM fitting procedures will not be adequate. To obtain the parameter estimates of ZIP regression models, $\hat{\beta}$ *and* $\hat{\gamma}$, the Newton –Raphson method or the method of Fisher scoring can be used. However, the method of scoring is more appropriate for ZIP regression because the second derivative $l(\mu,\pi;y)$ can be simplified by taking expectations.

3.2 The Method Of Fisher Scoring

Assuming that μ and π in (3.4) are not functionally related. The first and second derivatives of *l* with respect to β and γ are

$$\frac{\partial \ell}{\partial \beta_j} = \frac{\partial \ell(\mu,\pi)}{\partial \mu_i}\frac{\partial \mu_i}{\partial \beta_j}$$

$$= \left\{\sum_{i=1}^{n}\left\{I_{(y_i=0)}\left[\frac{-(1-\pi_i)\mu_i e^{-\mu_i}}{\pi_i+(1-\pi_i)e^{-\mu_i}}\right]\right\} + I_{(y_i>0)}\left[y_i-\mu_i\right]x_{ij}, j=1,2,....p;\right. \tag{3.5}$$

$$\frac{\partial \ell}{\partial \gamma_r} = \frac{\partial \ell(\mu,\pi)}{\partial \pi_i}\frac{\partial \pi_i}{\partial \gamma_r}$$

$$= \sum_{i=1}^{n}\left\{I_{(y_i=0)}\left[\frac{1-e^{-\mu_i}}{\pi_i+(1-\pi_i)e^{-\mu_i}}\right] - I_{(y_i>0)}\left[\frac{1}{1-\pi_i}\right]\right\}z_{ir}, \quad r=1,2,....,q; \tag{3.6}$$

and

$$\frac{\partial^2 \ell}{\partial \beta_j \partial \beta_k} = \sum_{i=1}^{n}\left\{I_{(y_i=0)}\left[\frac{-(1-\pi_i)\mu_i[(1-\mu_i)\pi_i+(1-\pi_i)e^{-\mu_i}]e^{-\mu_i}}{\left[\pi_i+(1-\pi_i)e^{\mu_i}\right]^2}\right] + I_{(y_i>0)}\left[-\mu_i\right]\right\}x_{ij}x_{ik}, \quad j,k=1,2,...., p; \tag{3.7}$$

$$\frac{\partial^2 \ell}{\partial \gamma_r \partial \gamma_s} = \sum_{i=1}^{n}\left\{I_{(y_i=0)}\left[\frac{-(1-e^{-\mu_i})^2}{[\pi_i+(1-\pi_i)e^{-\mu_i}]^2}\right] - I_{(y_i>0)}\left[\frac{1}{(1-\pi_i)^2}\right]\right\} z_{ir}z_{is}, \; r,s=1,2,....q; \tag{3.8}$$

$$\frac{\partial^2 \ell}{\partial \beta_j \partial \gamma_r} = \frac{\partial^2 \ell}{\partial \gamma_s \partial \beta_j} = \sum_{i=1}^{n}\left\{I_{(y_i=0)}\left[\frac{\mu e^{-\mu_i}}{[\pi_i+(1-\pi_i)e^{-\mu_i}]^2}\right]\right\}x_{ij}z_{ir.} \tag{3.9}$$

Using the fact that

$$E[I_{(y_i=0)}] = P(Y_i=0) = \pi_i+(1-\pi_i)e^{-\mu_i}$$

and

$$E[I_{(y_i>0)}] = P(Y_i>0) = (1-\pi_i)(1-e^{-\mu_i})$$

$$-E\left(\frac{\partial^2 \ell}{\partial \beta_j \partial \beta_k}\right) = \sum_{i=1}^{n}\left\{\left[\frac{(1-\pi_i)\mu_i\left[(1-\mu_i)\pi_i+(1-\pi_i)e^{-\mu_i}\right]e^{-\mu_i}}{\pi_i+(1-\pi_i)e^{-\mu_i}}\right]+(1-\pi)\mu_i(1-e^{-\mu_i})\right\}x_{ij}x_{ik}, \tag{3.10}$$

$$-E\left(\frac{\partial^2 \ell}{\partial \gamma_r \partial \gamma_s}\right) = \sum_{i=1}^{n}\left\{\left[\frac{(1-e^{-\mu_i})^2}{\pi_i+(1-\pi_i)e^{-\mu_i}}\right]+\left[\frac{1-e^{-\mu_i}}{1-\pi_i}\right]\right\}z_{ir}z_{is}, \tag{3.11}$$

and

$$-E\left(\frac{\partial^2 \ell}{\partial \beta_j \partial \gamma_r}\right)=\sum_{i=1}^{n}\left\{\frac{-\mu_i e^{-\mu_i}}{\pi_i+(1-\pi_i)e^{-\mu_i}}\right\}x_{ij}z_{ir} \tag{3.12}$$

Hence the estimates of β and γ at the $(m+1)^{th}$ iteration, denotes by $\beta^{(m+1)}$ and $\gamma^{(m+1)}$, are given by

$$\begin{bmatrix}\beta^{(m+1)}\\ \gamma^{(m+1)}\end{bmatrix}=\begin{bmatrix}\beta^{(m)}\\ \gamma^{(m)}\end{bmatrix}+\left[I^{(m)}(\beta,\gamma)\right]^{-1}S^{(m)}(\beta,\gamma),$$

Where the score vector and the expected information matrix, respectively evaluated at $\beta=\beta^{(m+1)}$ and $\gamma=\gamma^{(m+1)}$ are as fallows.

$$S(\beta,\gamma)=\begin{bmatrix}S_\beta(\beta,\gamma)\\ S_\gamma(\beta,\gamma)\end{bmatrix}=\begin{bmatrix}\frac{\partial \ell(\mu,\pi)}{\partial \beta}\\ \frac{\partial \ell(\mu,\pi)}{\partial \gamma}\end{bmatrix}$$

And the expected information matrix $I(\beta,\gamma)$

$$I(\beta,\gamma)=\begin{bmatrix}I_{\beta\beta}(\beta,\gamma) & I_{\beta\gamma}(\beta,\gamma)\\ I_{\gamma\beta}(\beta,\gamma) & I_{\gamma\gamma}(\beta,\gamma)\end{bmatrix}, \tag{3.14}$$

Where the elements $I_{\beta\beta}, I_{\beta\gamma}=I_{\gamma\beta}{}^{T}\ and\ I_{\gamma\gamma}\ are, respectively,$

$$-E\left(\frac{\partial^2 \ell(\mu,\pi)}{\partial\beta\partial\beta^T}\right),-E\left(\frac{\partial^2 \ell(\mu,\pi)}{\partial\beta\partial\gamma}\right), and -E\left(\frac{\partial^2 \ell(\mu,\pi)}{\partial\gamma\partial\gamma^T}\right).$$

Denote the inverted expected Fisher information matrix $I(\beta,\gamma) as K(\beta,\gamma),$

$$K(\beta,\gamma)=\begin{pmatrix}I_{\beta\beta}(\beta,\gamma) & I_{\beta\gamma}(\beta,\gamma)\\ I_{\gamma\beta}(\beta,\gamma) & I_{\gamma\gamma}(\beta,\gamma)\end{pmatrix}=\begin{bmatrix}K_{\beta\beta}(\beta,\gamma) & K_{\beta\gamma}(\beta,\gamma)\\ K_{\gamma\beta}(\beta,\gamma) & K_{\gamma\gamma}(\beta,\gamma)\end{bmatrix} \tag{3.15}$$

Note that Var($\hat{\beta}$) $and\ Var(\hat{\gamma})$ are obtained from $I^{-1}(\beta,\gamma),\ in\ (3.13)$ using inverse of partitioned matrix (Searle,1996) as follow

$$K^{\beta\beta}=Var(\hat{\beta})=\left[I_{\beta\beta}(\beta,\gamma)-I_{\beta\gamma}(\beta,\gamma)I_{\gamma\gamma}{}^{-1}(\beta,\gamma)I_{\gamma\beta}(\beta,\gamma)\right]^{-1} \tag{3.16}$$

$$K^{\gamma\gamma}=Var(\hat{\gamma})=\left[I_{\gamma\gamma}(\beta,\gamma)-I_{\gamma\beta}(\beta,\gamma)I_{\beta\beta}{}^{-1}(\beta,\gamma)I_{\beta\gamma}(\beta,\gamma)\right]^{-1} \tag{3.17}$$

With good stating values $\beta^0\ and\ \gamma^0\ \ and\ hence\ \mu^{(0)},\pi^{(0)}$ the iterative scheme converges in a few step, convergence is obtained with a stopping rule, such as $\left|\ell^{(m+1)}-\ell^m\right|\le e,\ \ where\ \ell^{(m+1)}\ and\ \ \ell^{(m)}\ are\ the$ $\log-likelihood, \ell(\mu,\pi;y)$ evaluated using the estimates of $\mu\ and\ \pi$ from the m and m+1 iterations, respectively. The asymptotic variance-covariance matrix for $\left(\hat{\beta},\hat{\gamma}\right)$ is automatically provided at the final iteration.

3.3. Maximum Likelihood Estimation for ZIP Models with no Covariates

Based on the log-likelihood function, the maximum likelihood estimates μ and π are the roots of the equations $\frac{\partial \ell(\mu,\pi)}{\partial \mu}=0\ and\ \frac{\partial \ell(\mu,\pi)}{\partial \pi}=0.$ here we have

$$\frac{\partial \ell(\mu,\pi)}{\partial \mu}=\frac{-\eta_0(1-\pi)e^{-\mu}}{\left[\pi+(1-\pi)e^{-\mu}\right]}-\sum_{j=1}^{J}\eta_j+\frac{\sum_{j=1}^{J}(\eta_j\times j)}{\mu}, \tag{3.12}$$

and

$$\frac{\partial \ell(\mu,\pi)}{\partial \pi}=\frac{\eta_0(1-e^{-\mu})}{\left[\pi+(1-\pi)e^{-\mu}\right]}-\frac{\sum_{j=1}^{J}(\eta_j)}{1-\mu}, \tag{3.13}$$

Setting each of these equal to zero gives

$$\frac{\eta_0\left(1-\hat{\pi}\right)e^{-\hat{\mu}}}{\left[\hat{\pi}+\left(1-\hat{\pi}\right)e^{-\hat{\mu}}\right]}+\sum_{j=1}^{J}\eta_j=\frac{\sum_{j=1}^{J}\left(\eta_j\times j\right)}{\hat{\mu}}$$

and

$$\frac{\eta_0\left(1-\hat{\pi}\right)}{\left[\hat{\pi}+\left(1-\hat{\pi}\right)e^{-\hat{\mu}}\right]}=\frac{\sum_{j=1}^{J}\eta_j}{\left(1-e^{\hat{\mu}}\right)}$$

Substituting (3.15) in to (3.14) gives

$$\frac{e^{-\hat{\mu}}\sum_{j=1}^{n}\eta_j}{\left(1-e^{-\hat{\mu}}\right)}+\sum_{j=1}^{n}\eta_j=\frac{\sum_{j=1}^{n}\left(\eta_j\times j\right)}{\hat{\mu}}$$

$$\sum_{j=1}^{n}\eta_j\left\{\frac{e^{-\hat{\mu}}+1-e^{-\hat{\mu}}}{\left(1-e^{-\hat{\mu}}\right)}\right\}=\frac{\sum_{j=1}^{n}\left(\eta_j\times j\right)}{\hat{\mu}}$$

$$\frac{\hat{\mu}}{\left(1-e^{-\hat{\mu}}\right)}=\frac{\sum_{j=1}^{n}\left(\eta_j\times j\right)}{\sum_{j=1}^{n}\eta_j}$$

$$\hat{\mu}=\frac{\left(1-e^{-\hat{\mu}}\right)\sum_{j=1}^{n}\left(\eta_j\times j\right)}{\sum_{j=1}^{n}\eta_j} \tag{3.16}$$

Note that this does not depend on ω or,η_0 from (3.15)

$$\eta_0\left(1-e^{-\hat{\mu}}\right)\left(1-\hat{\pi}\right)=\left[\hat{\pi}+\left(1-\hat{\pi}\right)e^{-\hat{\mu}}\right]\sum_{j=1}^{n}\eta_j$$

$$\eta_0\left(1-e^{-\hat{\mu}}\right)-\hat{\pi}\eta_0\left(1-e^{-\hat{\mu}}\right)=\hat{\pi}\sum_{j=1}^{n}\eta_j+\left(1-\hat{\pi}\right)e^{-\hat{\mu}}\sum_{j=1}^{n}\eta_j$$

$$\hat{\pi}\sum_{j=1}^{n}\eta_j-\hat{\pi}e^{-\hat{\mu}}\sum_{j=1}^{n}\eta_j+\hat{\pi}\eta_0\left(1-e^{-\hat{\mu}}\right)=\eta_0\left(1-e^{-\hat{\mu}}\right)-\ e^{-\hat{\mu}}\sum_{j=1}^{n}\eta_j$$

$$\hat{\pi}=\frac{\eta_0-\left(\eta_0+\sum_{j=1}^{n}\eta_j\right)e^{-\hat{\mu}}}{\left(\sum_{j=1}^{n}\eta_j+\eta_0\right)-\left(\sum_{j=1}^{n}\eta_j+\eta_0\right)e^{-\hat{\mu}}}$$

$$\hat{\pi}=\frac{\eta_0-\eta e^{-\hat{\mu}}}{\eta\left(1-e^{-\hat{\mu}}\right)} \tag{3.17}$$

4. TEST STATISTIC FOR ZERO-INFLATION PROPOSED IN ZIP MODELS

Within the family of regression models, testing if a regression models is adequate corresponding to testing

$$H_0 : \pi = 0;\ H_1 : \pi > 0,$$

Here π is taken as constant.

There are a number of test statistics proposed for testing the hypothesis including Wald, score test, likelihood ratio test, chi-square test, test based on a confidence interval of probability zero-inflated counts and Cochran test. Most of the tests mentioned were derived based on a single homogeneous sample, i.e. μ and π are not depend upon covariates or constant. The expressions of the tests and their sampling distributions are summarized in following subsections.

4.1. Likelihood ratio test

Let $\hat{l}_r$ denote the value of the log-likelihood function evaluated at the restricted maximum likelihood estimates (for instance, the Poisson model), and $\hat{l}_u$ the value of the log-likelihood function evaluated at the unrestricted maximum likelihood estimates (for instance, the zero-inflated Poisson modell), and the likelihood ratio test based on the ratio of two log-likelihood functions can be written as

$$LRT = -2(\hat{l}_r - \hat{l}_u)$$

$$LRT = -2\left\{\sum_{i=1}^{n}\left\{\{y_i \ln(\lambda_i) - \lambda_i - \ln(y_i)\} - \eta_0 \ln[\pi + (1-\pi)e^{-\hat{\mu}}] - \sum_{y=1}^{J}\eta_y \ln\left[(1-\pi)\frac{e^{-\hat{\mu}}\hat{\mu}^y}{y!}\right]\right\}\right\}$$

$$= 2\left\{\eta_0 \ln\left(\frac{\eta_0}{\eta}\right) + (\eta - \eta_0)\left(\ln\left(\frac{\bar{y}}{\hat{\mu}}\right) - \hat{\mu}\right) + \eta\bar{y}(\ln\hat{\mu} + 1 - \ln\bar{y})\right\}$$

Where $\bar{y}$ is the mean of the observations and $\hat{\mu}$ is the estimated positive mean counts. This test statistic is approximately fallows chi-square distribution with one degree of freedom under the null hypothesis.

4.2. Score test

A score test for the hypothesis is proposed by Van den broek (1995). The test is derived based on the log-likelihood to obtain the ratio of the score vector, $\begin{bmatrix} \frac{\partial \ell(\mu,\pi)}{\partial \mu} \\ \frac{\partial \ell(\mu,\pi)}{\partial \pi} \end{bmatrix}$ and minus expected information,

$$\begin{bmatrix} -E\left(\frac{\partial^2 \ell(\mu,\pi)}{\partial \mu^2}\right) & -E\left(\frac{\partial^2 \ell(\mu,\pi)}{\partial \mu \partial \pi}\right) \\ -E\left(\frac{\partial^2 \ell(\mu,\pi)}{\partial \pi \partial \mu}\right) & -E\left(\frac{\partial^2 \ell(\mu,\pi)}{\partial \pi^2}\right) \end{bmatrix}, evaluated\ at\ \pi = 0$$

Or under H_0 true. Using mathematical algebra, the score statistic is defined by

$$S_\omega = \frac{\left(\eta_0 - \eta e^{-\hat{\mu}_0}\right)^2}{\eta e^{-\hat{\mu}_0}(1 - e^{-\hat{\mu}_0}) - \eta\bar{y}e^{-2\hat{\mu}_0}},$$

$\hat{\mu}_o$ is the estimate of the Poisson parameter under the null hypothesis. This statistic will have an asymptotic chi-square distribution with one degree of freedom under the null hypothesis.

4.3. Chi-square test

The chi-square statistic χ^2 is used to test if a sample of data came from a population with a specific distribution. The χ^2 is commonly defined by

$$\chi^2_\omega = \sum_{k=1}^{c}\frac{(O_k - E_k)^2}{E_k}$$

Where c denotes the number of classes(categories) decided for a given data set, O_k and E_k are observed frequencies and expected frequencies under the null hypothesis of the k^{th} class, respectively. When the null hypothesis is valid, χ^2_ω fallows an asymptotic chi-square distribution on (c-1) d.f.

4.4. Test based on a confidence interval of probability zero-inflated counts

It is possible to derive a test based on asymptotic normality of the estimate of the parameters. Following the statistical properties of ZIP,

$$E(\overline{Y}) = E(Y) = (1-\pi)\mu = \lambda$$

$$Var(\overline{Y}) = \frac{1}{n}Var(Y) = \frac{1}{n}\left\{\frac{(1-\pi)\lambda + \pi\lambda^2}{1-\pi}\right\}.$$

From the central limit theorem, the confidence interval can be written as

$$-Z_{\alpha/2} \le \frac{\overline{Y} - (1-\pi)\mu}{\sqrt{Var(\overline{Y})}} \le Z_{\alpha/2}$$

$$-Z_{\alpha/2}\sqrt{Var(\overline{Y})} \le \overline{Y} - (1-\pi)\mu \le Z_{\alpha/2}\sqrt{Var(\overline{Y})}$$

$$1 - \frac{\overline{Y} - Z_{\alpha/2}\sqrt{\frac{1}{\eta}\left\{\frac{(1-\pi)\lambda + \pi\lambda^2}{1-\pi}\right\}}}{\mu} \le \pi \le 1 - \frac{\overline{Y} + Z_{\alpha/2}\sqrt{\frac{1}{\eta}\left\{\frac{(1-\pi)\lambda + \pi\lambda^2}{1-\pi}\right\}}}{\mu}$$

$$1 - \frac{\overline{y} - Z_{\alpha/2}\sqrt{\{\overline{y} + \overline{y}[\hat{\mu} - \overline{y}]\}/\eta}}{\hat{\mu}} \le \omega \le 1 - \frac{\overline{y} + Z_{\alpha/2}\sqrt{\{\overline{y} + \overline{y}[\hat{\mu} - \overline{y}]\}/\eta}}{\hat{\mu}}.$$

Hence, a test based on a positive one sided confidence interval of probability zero-inflated counts can be obtained as

$$CI_\omega = 1 - \frac{\overline{y} + Z_\alpha\sqrt{\{\overline{y} + \overline{y}[\hat{\mu} - \overline{y}]\}/\eta}}{\hat{\mu}},$$

Where $\hat{\mu}$ is the estimated positive mean counts under H_1. The critical region of this test method is simply $CI_\omega > 0$. That is , when $CI_\omega > 0$, we reject the null hypothesis at α level of significance and the ZIP model should be used instead of a Poisson models.

4.5. The Cochran test

This test is also known as C test. The C test is used to test the assumption of constant variance of the residuals in the analysis of variance. This test is a ratio that relates the largest empirical variance of a particular treatment to the sum of the variances of the remaining treatments. The C test statistic for ZIP model was developed by Xie *et al.* (2001) can be written as follows:

Here if one random variable $X \sim \chi^2_1$ then $\sqrt{X} \sim N(0,1)$. The Cochran test is transformed their original chi-square into its corresponding standard normal form. Cochran proposed to test any single deviation (s_0-t_0) when t is estimated from the data

From $\chi^2 = \frac{P}{Var(P)} \Rightarrow C_\omega = \frac{P}{\sqrt{Var(P)}} \sim N(0,1)$

Here $P = \eta_0 - \eta e^{-\overline{y}} \quad Var(P) = Var(\eta_0 - \eta e^{-\overline{y}})$

$$C_\omega = \frac{(\eta_0 - \eta e^{-\overline{y}})}{[\eta e^{-\overline{y}}(1 - e^{-\overline{y}} - \overline{y}e^{-\overline{y}}]^{1/2}}$$

Under the null hypothesis, the test statistic C_ω is approximately normally distributed with zero mean and unit variance.

$$C^2_\omega = \frac{(\eta_0 - \eta e^{-\overline{y}})^2}{[\eta e^{-\overline{y}}(1 - e^{-\overline{y}} - \overline{y}e^{-\overline{y}}]}$$

Fallowing the relationship between N ~ (0,1) and chi-square , we found that C can be obtained as C has the form exactly the same as score test.

4.6 Wald test

The Wald test is a statistical test, typically used to test whether an effect exists or not. A Wald test can be used in a great variety of different models including models for dichotomous or binary variables and models for continuous variables. Under the aspect of the Wald statistical test, named after Abraham Wald, in this paper we will develop a Wald test for ZIP model with constant μ and π for testing the hypothesis

$$H_0 : \pi = 0 ; H_1 : \pi > 0,$$

Based on the basic idea of obtaining the Wald test is

$$W_\omega = \frac{\hat{\pi}^2}{Var(\hat{\pi})}$$

Where $\hat{\pi}$ is the maximum likelihood estimate of ω under the ZIP model.

$$\hat{\pi} = \frac{(\eta_0 - \eta e^{-\hat{\mu}})}{\eta(1-e^{-\hat{\mu}})}$$

$$Var(\hat{\pi}) = (I_{\pi\pi} - I_{\pi\mu} I_{\mu\mu}^{-1} I_{\mu\pi})^{-1}$$

Then

$$I_{\pi\pi} - I_{\pi\mu} I_{\mu\mu}{}^{-1} I_{\mu\pi} = \eta \left\{ \frac{\left(1-e^{-\hat{\mu}}\right)}{\left(1-\hat{\pi}\right)\left[\hat{\pi}+\left(1-\hat{\pi}\right)e^{-\hat{\mu}}\right]} - \frac{\hat{\mu}e^{-2\hat{\mu}}}{\left(1-\hat{\pi}\right)\left[\hat{\pi}+\left(1-\hat{\pi}\right)e^{-\hat{\mu}}\right]\left\{\left[\hat{\pi}+\left(1-\hat{\pi}\right)e^{-\hat{\mu}}\right]-\hat{\pi}\hat{\mu}e^{-\hat{\mu}}\right\}} \right\}$$

$$= \frac{\eta\left\{\left(1-e^{-\hat{\mu}}\right)\left[\hat{\pi}+\left(1-\hat{\pi}\right)e^{-\hat{\mu}}\right]-\hat{\pi}\hat{\mu}e^{-\hat{\mu}}\right\}\hat{\mu}e^{-2\hat{\mu}}}{\left(1-\hat{\pi}\right)\left[\hat{\pi}+\left(1-\hat{\pi}\right)e^{-\hat{\mu}}\right]\left\{\left[\hat{\pi}+\left(1-\hat{\pi}\right)e^{-\hat{\mu}}\right]-\hat{\pi}\hat{\mu}e^{-\hat{\mu}}\right\}} \tag{3.28}$$

Since

$$\hat{\pi}+\left(1-\hat{\pi}\right)e^{-\hat{\mu}} = \frac{\eta_0}{\eta} and \left(1-\pi\right) = \frac{\bar{y}}{\hat{\mu}}, \tag{3.28}$$

$$I_{\pi\pi} - I_{\pi\mu} I_{\mu\mu}{}^{-1} I_{\mu\pi} = \frac{\eta^2\hat{\mu}\left\{\left(1-e^{-\hat{\mu}}\right)\left[\eta_0-\left(\hat{\mu}-\bar{y}\right)\eta e^{-\hat{\mu}}\right]-\eta\hat{\mu}e^{-2\hat{\mu}}\right\}}{\eta_0\bar{y}\left(\eta_0-\left(\hat{\mu}-\bar{y}\right)\eta e^{-\hat{\mu}}\right)}. \tag{3.29}$$

Hence

$$Var\left(\hat{\pi}\right) = \left(I_{\pi\pi} - I_{\pi\mu} I_{\mu\mu}{}^{-1} I_{\mu\pi}\right)^{-1} = \frac{\eta_0\bar{y}\left[\eta_0-\eta e^{-\hat{\mu}}\left(\hat{\mu}-\bar{y}\right)\right]}{\eta^2\hat{\mu}\left\{\left(1-e^{-\hat{\mu}}\right)\left[\eta_0-\eta e^{-\hat{\mu}}\left(\hat{\mu}-\bar{y}\right)\right]-\eta\hat{\mu}e^{-2\hat{\mu}}\right\}} \tag{3.30}$$

The Wald test for ZIP model is given as

$$W_\omega = \frac{\left(\frac{(\eta_0-\eta e^{-\hat{\mu}})}{\eta(1-e^{-\hat{\mu}})}\right)^2}{\left(\frac{\eta_0\bar{y}[\eta_0-\eta e^{-\hat{\mu}}(\hat{\mu}-\bar{y})]}{\eta^2\hat{\mu}\{(1-e^{-\hat{\mu}})[\eta_0-\eta e^{-\hat{\mu}}(\hat{\mu}-\bar{y})]-\eta\hat{\mu}e^{-2\hat{\mu}}\}}\right)}$$

$$= \frac{(\eta_0-\eta e^{-\hat{\mu}})^2\hat{\mu}\{(1-e^{-\hat{\mu}})[\eta_0-\eta e^{-\hat{\mu}}(\hat{\mu}-\bar{y})]-\eta\hat{\mu}e^{-2\hat{\mu}}\}}{\eta_0\bar{y}(1-e^{-\hat{\mu}})^2[\eta_0-\eta e^{-\hat{\mu}}(\hat{\mu}-\bar{y})]}$$

5. SIMULATION STUDY

In this section we conduct a small simulation study using R in order to investigated the performance of the Wald test for zero-inflation in small and moderate sample sizes within the ZIP regression model. Here the Wald test compares the Poisson model to the ZIP model. Here the simulation study was considered to examine the sample size and power of the wald tests. The outcome variable was generated from ZIP regression. Samples of size n=50,100,150 and 200 for each sample size n we simulated 1000 set of responses are generated under Poisson model and ZIP model with parameters μ =2.00 and π=0.00, 0.25, 0.45. For each set of generated data, a ZIP model is fitted for calculating the Wald and other tests are followed by the powers of the tests. Results from the simulation study are presented in tables

We can see that for a fixed value of π, the power of the those tests increases when sample size n increases. Similarly pattern is also found for increasing value of π and fixed n. here we can see the that when the value of π increases, number of excess zeros is large. The power of those tests are also good for this situation. For example when the sample size is 200, and parameters are 2.00 and 0.45 for μ and π respectively, the powers are higher than 0.993. Additionally, these tests are all good for comparing between Poisson and ZIP models. It can it can be seen that wald test is good as the Cochran and confidence interval tests, but it is better than likelihood ratio test, score test and chi-square test. Thus, the Wald test can be an alternative test for comparing between Poisson and ZIP models.

Table-5.1: Power of Wald, LRT, Score, Chi-square, Cochran, Confidence interval test statistics without covariates based on 1000 samples from the ZIP model with at α=0.01, μ =2.00

Level of the tests		**α=0.01**					
		Wald	**LRT**	**Score**	**Chi square**	**Cochran**	**Confidence interval**
n=50	**π=0.00**	0.008	0.007	0.014	0.08	0.07	0.007
	π=0.25	0.244	0.151	0.152	0.148	0.138	0.257
	π=0.45	0.683	0.471	0.472	0.368	0.263	0.693
n=100	**π=0.00**	0.006	0.05	0.016	0.028	0.021	0.007
	π=0.25	0.397	0.276	0.288	0.212	0.198	0.418
	π=0.45	0.871	0.734	0.740	0.583	0.503	0.895
n=150	**π=0.00**	0.004	0.03	0.018	0.032	0.018	0.003
	π=0.25	0.602	0.596	0.782	0.797	0.679	0.713
	π=0.45	0.918	0.839	0.856	0.801	0.734	0.902
n=200	**π=0.00**	0.02	0.02	0.011	0.005	0.003	0.02
	π=0.25	0.842	0.795	0.800	0.851	0.783	0.840
	π=0.45	0.999	0.997	0.997	0.993	0.994	0.995

Level of the tests		**α=0.05**					
		Wald	**LRT**	**Score**	**Chi square**	**Cochran**	**Confidence interval**
n=50	**π=0.00**	0.023	0.019	0.047	0.053	0.048	0.028
	π=0.25	0.407	0.339	0.340	0.289	0.183	0.483
	π=0.45	0.804	0.680	0.685	0.765	0.673	0.821
n=100	**π=0.00**	0.027	0.030	0.068	0.018	0.134	0.038
	π=0.25	0.594	0.504	0.510	0.474	0.283	0.603
	π=0.45	0.931	0.888	0.884	0.784	0.694	0.927
n=150	**π=0.00**	0.020	0.023	0.052	0.092	0.021	0.039
	π=0.25	0.681	0.718	0.810	0.923	0.731	0.789
	π=0.45	0.952	0.902	0.995	0.832	0.699	0.892
n=200	**π=0.00**	0.019	0.019	0.060	0.120	0.083	0.023
	π=0.25	0.934	0.918	0.919	0.958	0.831	0.913
	π=0.45	1.00	1.00	1.00	0.999	0.987	1.00

5.1 Example

The data in this study is based on 2015 road traffic accidents in Visakhapatnam which is obtained from the government of India, ministry of road transport & highways transport research wing new Delhi and also The Data collected from Visakhapatnam City according to FIR reports available at Traffic control room. The data provide information on road traffic accidents that occur within 365 consecutive days from January to December 2015. The data set is shown in the table. The observed and predicted probability for each count (from fitted models) for Poisson and ZIP model are also presented in table. There are 1438 observations, and the mean count is 0.36 with variance 0.58, the range of counts is from 0 to 6. The zero-fraction 1005/1438=0.6988 enable us to try the zero-inflated model. The observed percent for Poisson and ZIP are shown in figure 2. Here the Wald statistics is 89.734 with p value as $p<0.0001$ and remaining test statistics are indicating least evidence against the fit of the Poisson model to the data, Now we can use ZIP model to fit the data We can also use the AIC and BIC to present the additional evidence which is shown in table 4. This indicates that ZIP model is better predicted the data than the Poisson model since its value for the AIC and BIC are small. Also the figure 2 indicates that the ZIP model is a better choice than the Poisson model, since it can get better predicted percent.

Table-5.1: Observed and predicted probability from Poisson, ZIP model for the number of Road accidents death dataset

Count	Frequency	Observed probability	Predicted probability	
			Poisson	ZIP
0	1005	0.6989	0.6951	0.6989
1	387	0.2691	0.2528	0.2462
2	30	0.0209	0.0460	0.0480
3	9	0.0063	0.0056	0.0063
4	4	0.0028	0.0005	0.0006
5	0	0.0000	0.0000	0.0000
6	0	0.0000	0.0000	0.0000

Table-5.2: Model selection criteria for Poisson and ZIP regression models for the number of Road accidents death dataset

Selection Criteria	Models	
	Poisson	ZIP
Log-likelihood	884.115	**809.027**
AIC	1794.230	1664.054
BIC	1862.753	1785.287

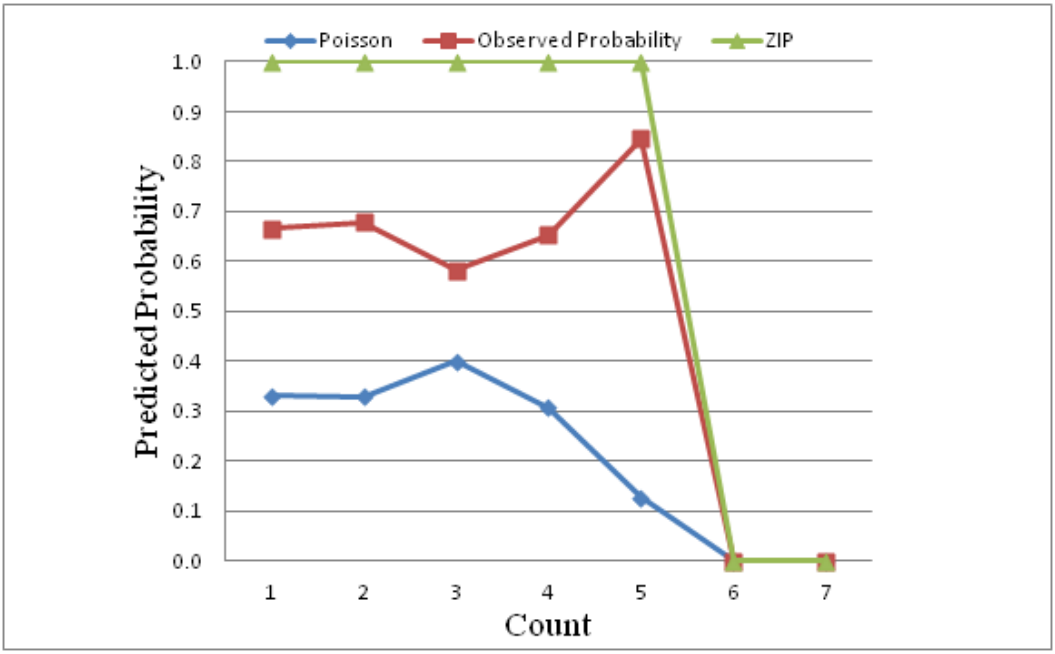


Figure-5.1: The Observed and predicted probability of Poisson and ZIP Models

6. DISCUSSION

Overdispersion is a common phenomenon in Poisson modeling, and this can be carried over to zero-inflated count data modeling. When zero-inflation exist in the count data, the ZIP model is frequently used. In this study, we have studied the properties of the score statistics, Cochran, Confidence interval, Chi-square, LRT, and Wald test statistics are examined via Monte Carlo simulations and application examples are given to illustrate our method.

We have also focused on the test statistic for testing the overdispersion used to compare the Poisson and ZIP models. Further, the properties of Wald, LRT and score tests examined and compared via a simulation study. The expression of the tests and their sampling distributions has been summarised. The results also indicated that the Wald test is preferable in terms of its small sample sizes. The Monte Carlo simulation indicates that for dataset that has small overdispersion parameter values the Poisson model is more appropriate while ZIP regression model is more appropriate for data that has high overdispersion values. Although Poisson and ZIP may be appropriate for different datasets. We can see this point from the example in this study, and the AIC, and BIC, test is used to explain the choice between Poisson and ZIP regression model.

7. REFERENCES

1. Agresti A. (1996), An Introduction to Categorical Data Analysis, wiley.
2. Breslow, N. (1990). Tests of hypotheses in over-dispersed Poisson regression and other quasi- likelihood models. J. Amer. Statist. Assoc. 85, 565-571.
3. Cameron A.C. and Trivedi P.K. (1990). Regression Based Test for Overdispersion in the Poisson Model. Journal of Applied Econometrics, 46:347- 364.
4. Cameron A.C. and Trivedi P.K. (1998). Regression Analysis of Count Data. Cambridge University Press.
5. Dean, C. B. (1992). Testing for overdispersion in Poisson and binomial regression models, J. Amer. Statist. Assoc. 87, 451-457.
6. Dejen Tesfaw.M and B.Muniswamy Power of Tests for Overdispersion Parameter in Negative Binomial Regression Model. IOSR Journal of Mathematics (IOSRJM) ISSN: 2278-5728 Volume 1, Issue 4 (July-Aug 2012), PP 29-36.
7. Jansakul, N. and Hinde, J. P. 2002. Score tests for zero-inflated Poisson models. Computational Statistics and Data Analysis. 40: 75-96.
8. Jansakul, N. and Hinde, J. P. 2009. Score tests for extra-zero models in zero inflated Negative Binomial models. Communications in statistics-simulation and computation. 38: 92-108.
9. Lambert, D. 1992. Zero-inflated Poisson regression with application to defects in manufacturing. Technometrics. 41(1): 29-38.
10. Mccullagh P. and Nelder J.A. (1989). Generalized Linear Models. Chapman and Hall, London.
11. Nelder, J. A. and Wedderburn, R. W. M. 1972. Generalized Linear Models. Journal of the Royal Statistical Society, Series A. 135(3): 370-384.
12. Ridout, M. S., Demetrio, C. G. B. and Hinde, J. P. 1998. Model for count data with many zeros. International Biometric Conference. 179-190.
13. Vuong Q.H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. Econometrica, 57:2:307-333.

Source of support: Nil, Conflict of interest: None Declared.