COMMUNICATIONS IN STATISTICS
Theory and Methods
Vol. 33, No. 1, pp. 47–64, 2004



# Score Test for Zero Inflated Generalized Poisson Regression Model

**Pushpa Lata Gupta,[1] Ramesh C. Gupta,[1,*] and Ram C. Tripathi[2]**

[1]Department of Mathematics and Statistics, University of Maine, Orono, Maine, USA
[2]Division of Management Sciences and Statistics, University of Texas, San Antonio, Texas, USA

## ABSTRACT

In certain applications involving count data, it is sometimes found that zeros are observed with a frequency significantly higher (lower) than predicted by the assumed model. Examples of such applications are cited in the literature from engineering, manufacturing, economics, public health, epidemiology, psychology, sociology, political science, agriculture, road safety, species abundance, use of recreational facilities, horticulture and criminology. In this article, a zero adjusted generalized Poisson distribution is studied and a score test is developed, with and without covariates, to determine whether such

*Correspondence: Ramesh C. Gupta, Department of Mathematics and Statistics, University of Maine, Orono, ME 04469-5752, USA; E-mail: rcgupta@maine.maine.edu.



DOI: 10.1081/STA-120026576 0361-0926 (Print); 1532-415X (Online)

an adjustment is necessary. Examples, with and without covariates, are provided to illustrate the results.

*Key Words:* Zero-inflated model; Generalized Poisson distribution; Score test; Covariate.

## 1. INTRODUCTION

Poisson model has been extensively used for the analysis of the count data. However, in Poisson model, mean variance relationship is quite restrictive in the presence of extra zeros. It underestimates the observed dispersion, which may be caused by extra zeros in the data. In the literature several authors have analyzed such data sets by the use of zero inflated Poisson, negative binomial or binomial models; see Bohning et al. (1999), Dietz and Bohning (2000), Gupta et al. (1996), Ridout et al. (2001), Van den Broek (1995) and the references therein.

The phenomenon of excess zeros (inflated) can arise as a result of clustering; distributions with clustering interpretations exhibit the feature that the proportion of observations in the zero class is greater than the estimate of the probability of zeros given by the assumed model. In an application in manufacturing, Lambert (1992) considered zero inflated Poisson regression models where she says, "One interpretation is that slight, unobserved changes in the environment cause the process to move randomly back and forth between perfect state in which the defects are extremely rare and an imperfect state in which the defects are possible but not inevitable". Mullahay (1997) has demonstrated that the unobserved hetrogeneity, commonly assumed to be the source of overdispersion in count data models, has predictable implications for the probability structures of such mixture models. In particular, the common observation of excess zeros implies that there is some unobserved hetrogeneity present in the data. This result has important implications, using count data, for predicting certain interesting parameters.

There are various models available in the literature to adjust for the hetrogeneity when the data are overdispersed, see for instance Dean and Lawless (1989) and Lawless (1987). Our interest, however, in this paper is to replace zero inflated Poisson distribution with the zero inflated generalized Poisson distribution since generalized Poisson distribution is a natural extension of the ordinary Poisson distribution in the non inflated situation. Thus, our purpose is to propose some alternatives, to the models already present in the literature, for zero inflated Poisson situation. The example presented is for illustration purposes of our model and is not meant to compare our procedure with those of the existing ones.

In order to adjust for extra (fewer) zeros, Gupta et al. (1995, 1996) studied zero adjusted count data models. More specifically, they proposed a zero adjusted generalized Poisson distribution and studied the relative error incurred by ignoring the adjustment. They also provided examples where the zero adjusted generalized Poisson distribution fits very well. In this paper, we study the zero inflated generalized Poisson regression model and develop a score test to determine whether an adjustment for inflation is necessary. Examples are provided, with and without the covariates, to illustrate the procedure. The organization of the paper is as follows: Section 2 contains the model and a procedure to develop the score test. In Sec. 3, we develop the score test without the covariates and in Sec. 4, a score test is developed taking into account the covariates. In Sec. 5, an example is provided to illustrate the procedure. This example deals with the data on the number of roots produced by micro-propogated shoots of the columnar apple cultivar Trajan and is taken from Ridout et al. (2001). The data is analyzed using the inflated generalized Poisson distribution, with and without covariates. In both cases, it is observed that the data fits the model well. Score tests are then used to establish the importance of the extra parameter of the generalized Poisson as well as the inflation parameter. Finally, in Sec. 6, some conclusions and recommendations are provided to analyze such data sets more accurately.

## 2. THE MODEL

Suppose the assumed model is represented by a discrete random variable W whose mass is concentrated on the non-negative integers. Suppose the zero class is observed with frequency significantly higher than predicted by the assumed model. Then the observed (inflated) random variable $Y$ can be described as

$$\begin{aligned} P(Y=0) &= \phi + (1-\phi)P(W=0) \\ P(Y=j) &= (1-\phi)P(W=j), \quad j = 1, 2, 3, \ldots \end{aligned} \tag{2.1}$$

where $0 \leq \phi < 1$. The model incoroporated extra zeros than given by the original model. Such a distribution can be regarded as a mixture of two distributions, one of which is degenerate at zero, see Johnson et al. (1992, page 312). We are excluding the possibility $\phi = 1$, because in that case the entire mass is concentrated at 0.

The subject of this paper is the generalized Poisson distribution (GPD) whose probability mass function is given by

$$P(W = x) = \frac{(1 + \alpha x)^{x-1}}{x!} \frac{(\theta e^{-\alpha\theta})^x}{e^{\theta}}, \quad x = 0, 1, 2, \ldots \tag{2.2}$$

The parameter space is as follows:

(1) $\theta > 0$, $\alpha \geq 0$, $0 \leq \alpha\theta < 1$.
(2) $\theta > 0$, $\alpha \leq 0$, $\max(-1, -\theta/m) < \alpha\theta \leq 0$, where $m$ is the largest positive integer such that $1 + \alpha m > 0$, see Consul (1989) and Johnson et al. (1992, page 396).

For $\alpha = 0$, it reduces to the ordinary Poisson distribution. The above model is a special case of the modified power series distribution, see Gupta (1974). The distribution of $Y$ is given by

$$\begin{aligned} P(Y = 0) &= \phi + (1 - \phi)e^{-\theta} \\ P(Y = y) &= (1 - \phi)\frac{(1 + \alpha y)^{y-1}}{y!} \frac{(\theta e^{-\alpha\theta})^y}{e^{\theta}}, \quad y = 1, 2, 3, \ldots \end{aligned} \tag{2.3}$$

Gupta et al. (1995) have studied the inference about the parameters of the above model under the umbrella of modified power series distributions. The effects of such an adjustment on the failure rates and the survival functions have also been examined by Gupta et al. (1996).

Letting $\psi = \phi/(1 - \phi)$, the model (2.3) can be written as

$$\begin{aligned} P(Y = 0) &= \frac{1}{1 + \psi}(\psi + e^{-\theta}) \\ P(Y = y) &= \frac{1}{1 + \psi}\frac{(1 + \alpha y)^{y-1}}{y!} \frac{(\theta e^{-\alpha\theta})^y}{e^{\theta}}, \quad y = 1, 2, 3, \ldots \end{aligned} \tag{2.4}$$

Note that when $0 \leq \phi < 1$, $\psi \geq 0$.

The null hypothesis for testing GPD vs. inflated GPD is to test $\phi = 0$ or equivalently $\psi = 0$. In the subsequent sections, we shall develop score test for testing $\psi = 0$ using covariates and without covariates. Before proceeding further, we briefly outline the procedure for constructing a score test.

### 2.1. Construction of a Score Test

Let $L(\theta)$ be the log-likelihood based on a random sample $Y_1, Y_2, \ldots, Y_n$ from a distribution with probability density function $f(x, \theta)$, where $\theta = (\theta_1, \theta_2, \ldots, \theta_k)'$ is a vector of unknown parameters taking values in $R^k$. The score vector $U(\theta)$ has components $U_i(\theta)$, $i = 1, 2, \ldots, k$, which are the partial derivatives of the log-likelihood with respect to the respective $\theta_i$. The Fisher information matrix $I(\theta)$ has entries $I_{ij}(\theta)$, for $i, j = 1, 2, \ldots, k$, where

$$I_{ij}(\theta) = E\left(-\frac{\partial^2 L(\theta)}{\partial\theta_i \partial\theta_j}\right).$$

It is well known that, under some mild conditions, $U(\theta)$ is asymptotically normal with mean vector $0'$ and covariance matrix $I(\theta)$. The statistic $U'(\theta_0)[I(\theta_0)]^{-1}U(\theta_0)$, which has asymptotically a chi-square distribution with $k$ degrees of freedom can be used to test the hypothesis $H_0 : \theta = \theta_0$. We are interested in testing a subset of $\theta_i$'s. The vector $\theta$ may be partitioned as $\theta = (\theta_1', \theta_2')'$, where $\theta_1$ is a $(p \times 1)$ vector and $\theta_2$ is a $(k - p \times 1)$ vector. The score vectors $U(\theta)$, $I(\theta)$ and $I^{-1}(\theta)$ are partitioned in a corresponding way as

$$U(\theta) = ([U_1(\theta)]', [U_2(\theta)]')'$$

$$\begin{bmatrix} I_{11}(\theta) & I_{12}(\theta) \\ I_{21}(\theta) & I_{22}(\theta) \end{bmatrix}$$

and

$$I^{-1}(\theta) = \begin{bmatrix} I^{11}(\theta) & I^{12}(\theta) \\ I^{21}(\theta) & I^{22}(\theta) \end{bmatrix}.$$

For a given $\theta_1 = \theta_{10}$, let $\tilde{\theta}_2(\theta_{10})$ be the MLE of $\theta_2$, which is obtained by maximizing $L((\theta_{10}', \theta_2')')$. Then, with $\tilde{\theta} = (\theta_{10}', \tilde{\theta}_2'(\theta_{10}))'$, the statistic $[U_1(\tilde{\theta})]'(I^{11}(\tilde{\theta}))[U_1(\tilde{\theta})]$, which has asymptotically a chi-square distribution with $p$ degrees of freedom, can be used to test the hypothesis $H_0 : \theta_1 = \theta_{10}$.

## 3. SCORE TEST WITHOUT COVARIATES

Based on a random sample $Y_1, Y_2, \ldots, Y_n$ from (2.4), the log-likelihood can be written as

$$\begin{aligned}
&\ln L(\psi, \theta, \alpha; y_1, y_2, \ldots, y_n) \\
&\quad = -n\ln(1+\psi) + \sum_i I_{\{y_i=0\}} \ln(\psi + e^{-\theta}) \\
&\qquad + \sum_i I_{\{y_i>0\}}(y_i - 1)\ln(1+\alpha y_i) + \sum I_{\{y_i>0\}} y_i(\ln\theta - \alpha\theta) \\
&\qquad - \sum_i I_{\{y_i>0\}} \ln y_i! - \sum_i I_{\{y_i>0\}}\theta.
\end{aligned} \tag{3.1}$$

The above log-likelihood function yields the following scores

$$U_1(\psi,\theta,\alpha) = \frac{\partial}{\partial\psi}\ln L = \frac{-n}{1+\psi} + \frac{n_0}{\psi + e^{-\theta}}. \tag{3.2}$$

$$U_2(\psi,\theta,\alpha) = \frac{\partial}{\partial\theta}\ln L = \frac{n_0\psi}{\psi + e^{-\theta}} + n\left(\left(\frac{1}{\theta} - \alpha\right)\bar{y} - 1\right). \tag{3.3}$$

$$U_3(\psi,\theta,\alpha) = \frac{\partial}{\partial\alpha}\ln L = \sum_{i=1}^{n} \frac{y_i(y_i - 1)}{1+\alpha y_i} - \theta n\bar{y}. \tag{3.4}$$

Here $n_0$ = number of zeros in the sample.

The Fisher information matrix $I(\psi, \theta, \alpha)$ is given by

$$I(\psi,\theta,\alpha) = \begin{bmatrix} \dfrac{n(1-e^{-\theta})}{(1+\psi)^2(\psi+e^{-\theta})} & \dfrac{-ne^{-\theta}}{(\psi+e^{-\theta})(1+\psi)} & 0 \\ \dfrac{-ne^{-\theta}}{(\psi+e^{-\theta})(1+\psi)} & \dfrac{-n\psi e^{-\theta}}{(1+\psi)(\psi+e^{-\theta})} + \dfrac{n}{\theta(1+\psi)(1-\alpha\theta)} & \dfrac{n\theta}{(1+\psi)(1-\alpha\theta)} \\ 0 & \dfrac{n\theta}{(1+\psi)(1-\alpha\theta)} & E\left(\sum_{i=1}^{n} \dfrac{Y_i^2(Y_i-1)}{(1+\alpha Y_i)^2}\right) \end{bmatrix}$$

The score vector is

$$\left(\frac{\partial}{\partial\psi}\ln L, \frac{\partial}{\partial\theta}\ln L, \frac{\partial}{\partial\alpha}\ln L\right)^T \Bigg|_{(0,\hat{\theta},\hat{\alpha})} = (n_0 e^{\hat{\theta}} - n, 0, 0)^T,$$

where $\hat{\theta}$ and $\hat{\alpha}$ are given by

$$\hat{\theta} = \frac{\overline{y}}{1 + \hat{\alpha}\overline{y}} \tag{3.5}$$

and

$$\sum_{i=1}^{n} \frac{y_i(y_i - 1)}{1 + \hat{\alpha} y_i} - \hat{\theta} n \overline{y} = 0. \tag{3.6}$$

Writing $J = [J_{ij}] = [I(\psi, \theta, \alpha)]|_{(0,\hat{\theta},\hat{\alpha})}$, the test statistic becomes

$$T = [a, 0, 0] J^{-1} \begin{bmatrix} a \\ 0 \\ 0 \end{bmatrix},$$

where $a = n_0 e^{\hat{\theta}} - n$. Thus $T$ is given by

$$T = a^2 J^{11} = a^2 \left[ \frac{J_{22}J_{33} - J_{23}^2}{J_{11}(J_{22}J_{33} - J_{23}^2) - J_{12}^2 J_{33}} \right],$$

where $J^{11}$ is the cofactor of $J_{11.}$

Here

$$J_{11} = n(e^{\hat{\theta}} - 1), \quad J_{12} = -n, \quad J_{22} = \frac{n}{\hat{\theta}(1 - \hat{\alpha}\hat{\theta})}, \quad J_{13} = 0,$$

$$J_{33} = E\left( \sum_{i=1}^{n} \frac{(Y_i - 1) Y_i^2}{(\alpha Y_i + 1)^2} \right) \quad \text{and} \quad J_{23} = \frac{n\hat{\theta}}{1 - \hat{\alpha}\hat{\theta}}.$$

Using the above values, the test statistic becomes

$$T = \frac{(n_0 e^{\hat{\theta}} - n)^2}{n(e^{\hat{\theta}} - 1) - \frac{n^2 J_{33}}{\frac{n}{\hat{\theta}(1-\hat{\alpha}\hat{\theta})} J_{33} - \left( \frac{n\hat{\theta}}{1-\hat{\alpha}\hat{\theta}} \right)^2}}. \tag{3.7}$$

**Remark 1.** The analytic expression for $J_{33}$ is not feasable. For large $n$, for computational purposes, $J_{33}$ is evaluated by ignoring the expectation operator.

## 4. SCORE TEST WITH COVARIATES

Let us now write the model (2.4) as

$$\begin{aligned} P(Y_i = 0) &= \frac{1}{1+\psi}(\psi + e^{-\theta_i}) \\ P(Y_i = y_i) &= \frac{1}{1+\psi}\frac{(1+\alpha y_i)^{y_i-1}}{y_i!}\frac{(\theta_i e^{-\alpha\theta_i})^{y_i}}{e^{\theta_i}}, \\ &\quad y_i > 0, \quad i = 1, 2, \ldots, n. \end{aligned} \tag{4.1}$$

The log-likelihood is given by

$$\begin{aligned} \ln L(\psi, \theta, \alpha; y_1, y_2, \ldots, y_n) \\ = -n\ln(1+\psi) + \sum_i I_{\{y_i=0\}} \ln(\psi + e^{-\theta_i}) \\ + \sum_i I_{\{y_i>0\}}(y_i - 1)\ln(1+\alpha y_i) + \sum_i I_{\{y_i>0\}} y_i(\ln\theta_i - \alpha\theta_i) \\ - \sum_i I_{\{y_i>0\}} \ln y_i! - \sum_i I_{\{y_i>0\}}\theta_i. \end{aligned} \tag{4.2}$$

We now introduce the covariates by modeling

$$\ln\theta_i = \sum_{r=1}^{p} x_{ir}\beta_r, \quad i = 1, 2, \ldots, n. \tag{4.3}$$

The scores are given by

$$\frac{\partial}{\partial\psi}\ln L = \frac{-n}{1+\psi} + \sum_i I_{\{y_i=0\}}\frac{1}{\psi + e^{-\theta_i}}. \tag{4.4}$$

$$\begin{aligned} \frac{\partial}{\partial\beta_r}\ln L = -\sum_i I_{\{y_i=0\}}\frac{e^{-\theta_i}}{\psi + e^{-\theta_i}}\theta_i x_{ir} \\ + \sum_i I_{\{y_i>0\}}\{y_i - (\alpha y_i + 1)\theta_i\}x_{ir}, \quad r = 1, 2, \ldots, p. \end{aligned} \tag{4.5}$$

$$\frac{\partial}{\partial\alpha}\ln L = \sum_i I_{\{y_i>0\}}\left[\frac{y_i - 1}{1+\alpha y_i} - \theta_i\right]y_i. \tag{4.6}$$

The mixed derivatives are given by

$$\frac{\partial^2}{\partial\psi^2}\ln L = \frac{n}{(1+\psi)^2} - \sum_i I_{\{y_i=0\}}\frac{1}{(\psi+e^{-\theta_i})^2}. \tag{4.7}$$

$$\frac{\partial^2}{\partial\beta_r\partial\beta_s}\ln L = \sum_i I_{\{y_i=0\}}\frac{e^{-\theta_i}}{\psi+e^{-\theta_i}}\left[\frac{\psi\theta_i}{\psi+e^{-\theta_i}} - 1\right]\theta_i x_{ir}x_{is} - \sum_i I_{\{y_i>0\}}(1+\alpha y_i)\theta_i x_{ir}x_{is}, \quad r,s = 1,2,\ldots,p. \tag{4.8}$$

$$\frac{\partial^2}{\partial^2\alpha}\ln L = -\sum_i \frac{y_i^2(y_i-1)}{(1+\alpha y_i)^2}. \tag{4.9}$$

$$\frac{\partial^2}{\partial\beta_r\partial\alpha}\ln L = -\sum_i y_i\theta_i x_{ir}. \tag{4.10}$$

$$\frac{\partial^2}{\partial\psi\partial\alpha}\ln L = 0. \tag{4.11}$$

$$\frac{\partial^2}{\partial\psi\partial\beta_r}\ln L = \sum_i I_{\{y_i=0\}}\frac{e^{-\theta_i}\theta_i x_{ir}}{(\psi+e^{-\theta_i})^2}. \tag{4.12}$$

We also need

$$E(I_{\{Y_i=0\}}) = P(Y_i=0) = \frac{1}{1+\psi}(\psi+e^{-\theta_i}).$$

$$E(I_{\{Y_i>0\}}) = P(Y_i>0) = \frac{1-e^{-\theta_i}}{1+\psi}.$$

We now present the expected values of the mixed derivatives

$$-E\left(\frac{\partial^2\ln L}{\partial\beta_r\partial\beta_s}\right) = -\sum_i \frac{e^{-\theta_i}}{\psi+e^{-\theta_i}}\left(\frac{\psi\theta_i}{\psi+e^{-\theta_i}} - 1\right)\theta_i x_{ir}x_{is}E(I_{\{Y_i=0\}}) + \sum_i E[(1+\alpha Y_i)\theta_i x_{ir}x_{is}I_{\{Y_i>0\}}]$$





MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

$$= \sum_i \frac{e^{-\theta_i}\theta_i x_{ir} x_{is}}{1+\psi}\left(1 - \frac{\psi\theta_i}{\psi + e^{-\theta_i}}\right) + \sum_i \theta_i x_{ir} x_{is}\left(\frac{\alpha\mu_i}{1+\psi} + \frac{1-e^{-\theta_i}}{1+\psi}\right),$$
$$r, s = 1, 2, \ldots, p, \tag{4.13}$$

where

$$E(X_i) = \mu_i = \frac{\theta_i}{1-\theta_i\alpha}. \tag{4.14}$$

Also

$$-E\left(\frac{\partial^2 \ln L}{\partial^2\alpha}\right) = \sum_i E\left[\frac{Y_i^2(Y_i-1)}{(1+\alpha Y_i)^2}\right]. \tag{4.15}$$

$$-E\left(\frac{\partial^2 \ln L}{\partial\psi^2}\right) = \frac{1}{(1+\psi)^2}\sum_i \frac{1-e^{-\theta_1}}{\psi + e^{-\theta_i}}. \tag{4.16}$$

$$-E\left(\frac{\partial^2 \ln L}{\partial\beta_r\partial\alpha}\right) = \sum_i \theta_i x_{ir}\frac{\mu_i}{1+\psi}, \quad r = 1, 2, \ldots, p. \tag{4.17}$$

$$-E\left(\frac{\partial^2 \ln L}{\partial\beta_r\partial\psi}\right) = -\sum_i \frac{e^{-\theta_i}\theta_i x_{ir}}{(1+\psi)(\psi + e^{-\theta_i})}, \quad r = 1, 2, \ldots, p. \tag{4.18}$$

$$-E\left(\frac{\partial^2 \ln L}{\partial\psi\partial\alpha}\right) = 0. \tag{4.19}$$

The score vector under the null hypothesis is given by

$$U(\psi, \beta, \alpha)^T|_{(0,\hat{\beta}_1,\hat{\beta}_2,\ldots,\hat{\beta}_p,\hat{\alpha})}$$
$$= \left[\frac{\partial \ln L}{\partial\psi}, \frac{\partial \ln L}{\partial\beta_1}, \frac{\partial \ln L}{\partial\beta_2}, \ldots, \frac{\partial \ln L}{\partial\beta_p}, \frac{\partial \ln L}{\partial\alpha}\right]\Bigg|_{(0,\hat{\beta}_1,\hat{\beta}_2,\ldots,\hat{\beta}_p,\hat{\alpha})}$$
$$= [\hat{U}_1, 0, 0, \ldots 0],$$

where

$$\hat{U}_1 = \frac{\partial}{\partial\psi}\ln L|_{(0,\hat{\beta}_1,\hat{\beta}_2,\ldots,\hat{\beta}_p,\hat{\alpha})} = -n + \sum_i I_{\{y_i=0\}} e^{\hat{\theta}_i}.$$

Writing

$$J = [J_{ij}] = [I(\psi, \beta, \alpha)]|_{(0,\hat{\beta}_1,\hat{\beta}_2,\ldots,\hat{\beta}_p,\hat{\alpha})}$$

the test statistic can be written as

$$\begin{aligned} T &= [U(0,\hat{\beta},\hat{\alpha})]^T J^{-1}[U(0,\hat{\beta},\hat{\alpha})] \\ &= \hat{U}_1^{\,2} J^{11} = \hat{U}_1^{\,2} J^{11}(C_{11} - C_{12}C_{22}^{-1}C_{21})^{-1} \\ &= \frac{\hat{U}_1^{\,2}}{C_{11} - C_{12}C_{22}^{-1}C_{21}}, \end{aligned} \tag{4.20}$$

where $J^{11}$ is the cofactor of $J_{11}$ and

$$J = \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix}.$$

Here $C_{11}$ is $1 \times 1$, $C_{12}$ is $1 \times (p+1)$, $C_{21}$ is $(p+1) \times 1$ and $C_{22}$ is a $(p+1) \times (p+1)$ matrix. These are given by

$$C_{11} = \sum_{i=1}^{n} (e^{\hat{\theta}_i} - 1)$$

$$C_{22} = \begin{bmatrix} X^T diag(\hat{\mu})X & -X^T(\hat{\mu\theta}) \\ -(\hat{\mu\theta})^T X & K \end{bmatrix},$$

where

$$K = \sum_i E\left[\frac{Y_i^2(Y_i - 1)}{(1 + \alpha Y_i)^2}\right]$$

$$C_{12} = \left[(\hat{\theta})^T X \quad 0,\right]$$

where $X = [x_{ij}]_{n \times p}$, $diag(\mu) = [\mu_{ij}]_{n \times n,}$ $\mu_{ij} = \mu_i$, $i = j$, $\mu\theta = [\mu_i\theta_i]_{n \times 1} = [\mu_1\theta_1, \mu_2\theta_2, \ldots, \mu_n\theta_n]^T$.

So the test statistic becomes

$$T = \frac{[-n + \sum_{i=1}^n e^{\hat{\theta}_i} I_{\{Y_i=0\}}]^2}{\sum_{i=1}^n (e^{\hat{\theta}_i} - 1) - ((\hat{\theta})^T X) U^{-1} (X^T \hat{\theta})}, \tag{4.21}$$





Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

where

$$U = X^T diag(\hat{\mu})X - \frac{[X^T(\hat{\mu\theta})][(\hat{\mu\theta})^T X]}{K}.$$

**Remark 2.** As before the analytic expression for $K$ is not feasable. For numerical example, the expectation operator in the expression for $K$ can be ignored.

## 5. AN EXAMPLE

The following data are taken from Ridout et al. (2001). The data (Table 1) consist of the number of roots produced by 270 micro-propogated shoots of the columnar apple cultivar Trajan. The roots had been produced under an 8- or 16-h photoperiod in culture systems that utilized one of four different concentrations of the cytokinin BAP in culture medium. For illustration, we have merged the data on four concentrations into one group. Let Group I consist of the data produced under 8 h photo

***Table 1.***

| # of roots | Observed frequency (Group I) | Observed frequency (Group II) | Total |
|---|---|---|---|
| 0 | 2 | 62 | 64 |
| 1 | 3 | 7 | 10 |
| 2 | 6 | 7 | 13 |
| 3 | 7 | 8 | 15 |
| 4 | 13 | 8 | 21 |
| 5 | 12 | 6 | 18 |
| 6 | 14 | 10 | 24 |
| 7 | 17 | 4 | 21 |
| 8 | 21 | 2 | 23 |
| 9 | 14 | 7 | 21 |
| 10 | 13 | 4 | 17 |
| 11 | 10 | 2 | 12 |
| 12 | 2 | 3 | 5 |
| 13 | 2 | 0 | 2 |
| 14 | 3 | 0 | 3 |
| 15 | 0 | 0 | 0 |
| 16 | 0 | 0 | 0 |
| 17 | 1 | 0 | 1 |
| Total | 140 | 130 | 270 |





Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

period and Group II consist of the data produced under produced under 16 h photo period.

### 5.1. Analysis with Covariates

Here the total # of shoots is 270 of which 140 shoots belong to Group I and 130 shoots belong to Group II.

The link function is given by

$$\ln \theta_i = \beta_1 + x_{i2}\beta_2, \quad i = 1, 2, \ldots, 270, \tag{5.1}$$

where

$$\begin{aligned} x_{i2} &= 0 \quad \text{for Group I} \\ &= 1 \quad \text{for Group II.} \end{aligned}$$

This means that

$$\begin{aligned} \ln \theta_i &= \beta_1, \qquad i = 1, 2, \ldots, 140 \\ &= \beta_1 + \beta_2, \quad i = 141, \ldots, 270. \end{aligned}$$

Letting $e^{\beta_1} = a$ and $e^{\beta_2} = b$, the log-likelihood function given in Sec. 4, yields the following:

$$\frac{\partial}{\partial \psi} \ln L = \frac{-270}{1+\psi} + \frac{2e^a}{\psi e^a + 1} + \frac{62e^{ab}}{\psi e^{ab} + 1} \tag{5.2}$$

$$\begin{aligned} \frac{\partial}{\partial \beta_1} \ln L = 994 - \frac{2a}{\psi e^a + 1} - 994a\alpha - 138a + 372 \\ - \frac{62ab}{\psi e^{ab} + 1} - 372ab\alpha - 68ab \end{aligned} \tag{5.3}$$

$$\frac{\partial}{\partial \beta_2} \ln L = 372 - \frac{62ab}{\psi e^{ab} + 1} - 372ab\alpha - 68ab \tag{5.4}$$

$$\begin{aligned} \frac{\partial}{\partial \alpha} \ln L = \frac{26}{1+2\alpha} + \frac{90}{1+3\alpha} + \frac{252}{1+4\alpha} + \frac{360}{1+5\alpha} + \frac{720}{1+6\alpha} + \frac{882}{1+7\alpha} \\ + \frac{1288}{1+8\alpha} + \frac{1512}{1+9\alpha} + \frac{1530}{1+10\alpha} + \frac{1320}{1+11\alpha} + \frac{660}{1+12\alpha} + \frac{312}{1+13\alpha} \\ + \frac{546}{1+14\alpha} + \frac{272}{1+17\alpha} - 994a - 372ab. \end{aligned} \tag{5.5}$$

Solving the likelihood equations obtained by setting the above derivatives equal to zero, we get

$$\hat{\psi} = .295, \quad \hat{\beta}_1 = \ln a = 1.738, \quad \hat{\beta}_2 = \ln b = -.252, \quad \hat{\alpha} = .037.$$

The asymptotic 95% confidence intervals for $\psi$ and $\alpha$ are given by

$$\text{C.I. for } \psi : (.210, .381)$$

$$\text{C.I. for } \alpha : (.020, .054).$$

Merging the last three classes, the observed and the expected frequencies are given below in Table 2.

The Pearson chi-square statistic value gives 20.96 which is quite close to the 5% chi-square table value with 11 degrees of freedom. This indicates a reasonable fit of the data with the inflated generalized Poisson distribution (IGPD).

Assuming now that the data follow a IGPD, we developed the score tests for testing the hypotheses $H_{01} : \alpha = 0$ and $H_{02} : \psi = 0$.

***Table 2.***

| # of roots | Observed frequency | Expected frequency |
|---|---|---|
| 0 | 64 | 63.122 |
| 1 | 10 | 6.226 |
| 2 | 13 | 13.322 |
| 3 | 15 | 20.521 |
| 4 | 21 | 25.530 |
| 5 | 18 | 27.292 |
| 6 | 24 | 26.050 |
| 7 | 21 | 22.778 |
| 8 | 23 | 18.579 |
| 9 | 21 | 14.323 |
| 10 | 17 | 10.540 |
| 11 | 12 | 7.459 |
| 12 | 5 | 5.106 |
| 13 | 2 | 3.396 |
| 14 | 3 | 2.202 |
| ≥15 | 1 | 2.794 |

5.1.1. Score Test for α (Inflated Poisson vs. IGPD)

The likelihood equations are solved under $\alpha = 0$ and the estimates are

$$\hat{a} = 7.203, \quad \hat{b} = .750, \quad \hat{\psi} = .305.$$

The score statistic is $10.297 > \chi^2_{.05,1} = 3.841$.

Hence the inflated Poisson distribution is rejected in favor of the IGPD. This establishes the justification for using the inflated generalized Poisson distribution instead of the inflated Poisson distribution. Notice that the confidence interval, obtained earlier, for α also gives the same message.

**Remark 3.** The general expression for the score test statistic can be derived in exactly the same way as for testing $\psi = 0$.

5.1.2. Score Test for ψ (GPD vs. IGPD)

The likelihood equations are solved under $\psi = 0$ and the estimates are

$$\hat{a} = 3.304, \quad \hat{b} = .592, \quad \hat{\alpha} = .162.$$

The score statistic is $24.911 > \chi^2_{.05,1} = 3.841$.

Hence the GPD is rejected in favor of the IGPD. Notice that the confidence interval, obtained earlier, for ψ also gives the same message.

### 5.2. Analysis without Covariates

In this case, we have three parameters ψ, θ, α. The log likelihood function, given in Sec. 3, yields

$$\frac{\partial}{\partial\psi}\ln L = \frac{-270}{1+\psi} + \frac{64e^a}{\psi e^a + 1}$$

$$\frac{\partial}{\partial\theta}\ln L = 1366 - \frac{64a}{\psi e^a + 1} - 1366a\alpha - 206a \tag{5.6}$$

$$\begin{aligned}\frac{\partial}{\partial\alpha}\ln L = {} & \frac{26}{1+2\alpha} + \frac{90}{1+3\alpha} + \frac{252}{1+4\alpha} + \frac{360}{1+5\alpha} + \frac{720}{1+6\alpha} + \frac{882}{1+7\alpha} \\ & + \frac{1288}{1+8\alpha} + \frac{1512}{1+9\alpha} + \frac{1530}{1+10\alpha} + \frac{1320}{1+11\alpha} + \frac{660}{1+12\alpha} \\ & + \frac{312}{1+13\alpha} + \frac{546}{1+14\alpha} + \frac{272}{1+17\alpha} - 1366a.\end{aligned} \tag{5.7}$$

Solving the likelihood equations obtained by setting the above derivatives equal to zero, we get

$$\hat{\psi} = .303, \quad \hat{\theta} = 5.146, \quad \hat{\alpha} = .043.$$

The asymptotic 95% confidence intervals for $\psi$ and $\alpha$ are given by

$$\text{C.I. for } \psi : (.216,\ .390)$$

$$\text{C.I. for } \alpha : (.021,\ .064)$$

Merging the last three classes, the observed and the expected frequencies are given below in Table 3.

The Pearson chi-square statistic value gives 17.946 while the 5% chi-square table value with 12 degrees of freedom is 21.026. This indicates a good fit of the data with the inflated generalized Poisson distribution (IGPD).

Assuming now that the data follow a IGPD, we develope the score tests for testing the hypotheses $H_{01} : \alpha = 0$ and $H_{02} : \psi = 0$.

***Table 3.***

| # of roots | Observed frequency | Expected frequency |
|---|---|---|
| 0 | 64 | 64.000 |
| 1 | 10 | 4.988 |
| 2 | 13 | 11.182 |
| 3 | 15 | 18.052 |
| 4 | 21 | 23.508 |
| 5 | 18 | 26.238 |
| 6 | 24 | 26.054 |
| 7 | 21 | 23.598 |
| 8 | 23 | 19.844 |
| 9 | 21 | 15.595 |
| 10 | 17 | 11.793 |
| 11 | 12 | 8.484 |
| 12 | 5 | 5.880 |
| 13 | 2 | 3.945 |
| 14 | 3 | 2.573 |
| ≥15 | 1 | 3.318 |





Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

5.2.1. Score Test for $\alpha$ (Inflated Poisson vs. IGPD)

The likelihood equations are solved under $\alpha = 0$ and the estimates are

$$\hat{\theta} = 6.622, \quad \hat{\psi} = .309.$$

The score statistic is $16.821 > \chi^2_{.05,1} = 3.841$.

Hence the inflated Poisson distribution is rejected in favor of the IGPD. Notice that the confidence interval, obtained earlier, for $\alpha$ also gives the same message.

5.2.2. Score Test for $\psi$ (GPD vs. IGPD)

The likelihood equations are solved under $\psi = 0$ and the estimates are

$$\hat{\theta} = 2.208, \quad \hat{\alpha} = .255.$$

The score statistic is $159.669 > \chi^2_{.05,1} = 3.841$.

Hence the GPD is rejected in favor of the IGPD. Notice that the confidence interval, obtained earlier, for $\psi$ also gives the same message.

In conclusion, we may state that the IGPD is a more appropriate model for the analysis of the above data.

## 6. CONCLUSION AND REMARK

The zero inflated generalized Poisson model studied in this paper is a generalization of the Poisson as well as the inflated Poisson distributions. It takes into account the extra zeros present in the data than those predicted by the model. Ignoring this adjustment may lead to erroneous conclusions in data analysis. The example presented shows that the inflated generalized Poisson distribution is a better alternative to the usual Poisson and even inflated Poisson distributions. It is, therefore, recommended that in order to obtain more accurate results, the model should be adjusted for the number of zeros.

## ACKNOWLEDGMENT

The authors are thankful to the referee for some useful suggestions which enhanced the presentation.

## REFERENCES

Bohning, D., Dietz, E., Schlattman, P., Mendonca, L., Kirchner, U. (1999). The zero inflated Poisson model and the decayed, missing and filled teeth in dental epidemiology. *J. Royal Stat. Soc. A* 162:195–209.

Consul, P. C. (1989). *Generalized Poisson Distributions*. New York: Marcel Dekker.

Dean, C., Lawless, J. F. (1989). Tests for detecting overdispersion in Poisson regression models. *J. Am. Stat. Assoc.* 84:467–472.

Dietz, E., Bohning, D. (2000). On estimation of Poisson parameter in zero-modified Poisson models. *Comput. Stat. Data Anal.* 34:441–459.

Gupta, R. C. (1974). Modified power series distribution and its applications. In: *Sankhya*. Series B, pp. 288–298.

Gupta, P. L., Gupta, R. C., Tripathi, R. C. (1995). Inflated modified power series distributions with applications. *Commun. Stat. Theory Meth.* 24:2355–2374.

Gupta, P. L., Gupta, R. C., Tripathi, R. C. (1996). Analysis of zero adjusted count data. *Comput. Stat. Data Anal.* 23:207–218.

Johnson, N. L., Kotz, S., Kemp, A. W. (1992). *Univariate Discrete Distributions*. New York: John Wiley and Sons.

Lambert, D. (1992). Zero inflated Poisson regression with an application to defects in manufacturing. *Technometrics* 34:1–14.

Lawless, J. F. (1987). Negative binomial and mixed Poisson regression. *Canadian J. Stat.* 15(3):209–225.

Mullahay, J. (1997). Heterogeneity, excess zeros, and the structure of count data models. *J. Appl. Econometrics* 12:337–350.

Ridout, M., Hinde, J., Demetrio, C. G. B. (2001). A score test for testing zero inflated Poisson regression model against zero inflated negative binomial alternatives. *Biometrics* 57:219–223.

Van den Broek, J. (1995). A score test for zero inflation in a Poisson distribution. *Biometrics* 51:731–743.