MATEMATIKA, Volume 27, Number 1, 21–29

Zero-Inflated Poisson Regression Models with Right Censored Count Data

[1]**Seyed Ehsan Saffari and** [2]**Robiah Adnan**
[1]Department of Mathematics, Universiti Teknologi Malaysia
81310 UTM Johor Bahru, Malaysia
e-mail: [1]ehsanreiki@yahoo.com, [2]robiaha@utm.my

Abstract A Poisson model typically is assumed for count data. However Poisson model is not suitable for data because of too many zeros. Furthermore, the response variable in such cases is censored for some values. In this paper, a zero-inflated Poisson regression model is introduced on censored data. In this model, we consider a response variable and one or more than one explanatory variables. The estimation of regression parameters using the maximum likelihood method is discussed and the goodness-of-fit for the regression model is examined. We study the effects of right censoring in terms of parameters estimation and their standard errors via simulation and an example.

Keywords Zero-Inflated Poisson regression; Censored data ; Maximum likelihood method; Simulation

2010 Mathematics Subject Classification 62J02

1 Introduction

There are many statistical applications, when the random variable Y which is the dependent variable represents counts. Count data can take two forms that is, simple counts and categorical data, depends on how the data arise. Simple counts can be the number of occurrences of flash floods in a month, observed for several years. While categorical data in which the count represent the number of items belonging to each of the several categories. Statistical methods such as least square and analysis of variance are designed to deal with continuous variables. Due to this, many studies dealing with count data and various distributions have been proposed for the response or dependent variable, like Poisson distribution, negative binomial distribution, generalized Poisson distribution.

A Poisson distribution is frequently assumed in order to analyze count data, which implies equality of the mean and the variance. But in practice, the observed variability often violates this theoretical assumption. It is often the case that the sample variance is greater than or less than the observed sample mean and it is classified as under- or over-dispersion, respectively [1]. Another type of over-dispersion related to Poisson distribution is that in such cases the number of zero counts are much greater than expected for the Poisson distribution. Ridout has discussed about some examples of data with too many zeros from various disciplines, [2]. Regular approaches cannot be applied directly for modeling count data when excess zeros exist. In this case, statistical approaches for modeling count data with many zeros than expected have been studied. Lambert derived the zero-inflated Poisson regression (ZIPR) model and its asymptotic properties of the ML estimator [3].

In many applications, count data are often censored from above (right) or below (left) a specific point or a combination of them (interval). Actually, censoring from above or below a specific point are special cases of interval censoring. Censoring must never be confused with truncation. In truncation, observations never result in values outside a given range.

A truncated sample can be thought as a sample with all values outside the bounds entirely omitted, not even the count to those omitted were kept. Whereas when the sample had been censored, a record noting that whether the lower or upper bound had been passed and the value of the bound are available. The case of variable threshold was considered by Caudill and Mixon [4]. They also highly recommended the use of censored negative binomial regression model, when the censored count data is over-dispersed. To analyze censored data with a constant censoring threshold, Terza [5] proposed the censored Poisson regression (CPR) model and obtained the ML estimator using the Newton-Raphson method. In practice, censored count data are too dispersed to use the CP model.

In this article, the main objective is to explain how we can use zero-inflated Poisson regression model in right censored data. In section 2, the zero-inflated Poisson regression model is defined and the likelihood function of zero inflated regression model in right censored data is formulated. In section 3, the parameter estimation is discussed using maximum likelihood method. In section 4, the goodness-of-fit for the regression model is examined and a test statistic for examining the dispersion of zero-inflated regression model in right censored data is proposed. A simulation for a censored zero-inflated Poisson regression model in terms of the parameter estimation, standard errors and goodness-of-fit statistic is conducted in section 5.

2 The Model

Let $\boldsymbol{Y}_i$ be a nonnegative integer-valued random variable and suppose $\boldsymbol{Y}_i = 0$ is observed with a frequency significantly higher than can be modeled by the usual model. Thus, the regression model is defined as

$$P(\boldsymbol{Y} = y_i|x_i, z_i) = \begin{cases} \varphi_i + (1-\varphi_i)f(0;\boldsymbol{\theta}_i), & y_i = 0, \\ (1-\varphi_i)f(y_i;\boldsymbol{\theta}_i), & y_i > 0, \end{cases} \tag{1}$$

where $f(y_i;\boldsymbol{\theta}_i), y_i = 0,1,2,\ldots$ is the pdf of $\boldsymbol{Y}_i$ and $0 < \varphi_i < 1$. Furthermore, the function $\varphi_i = \varphi_i(z_i)$ satisfy $logit(\varphi_i) = \log(\varphi_i[1-\varphi_i]^{-1}) = \sum_{j=1}^{m} z_{ij}\delta_j$ where $z_i = (z_{i1} = 1, z_{i2}, \ldots, z_{im})$ is the i-th row of covariate matrix $\boldsymbol{Z}$ and $\boldsymbol{\delta} = (\delta_1, \delta_2, \ldots, \delta_m)$ are unknown m-dimensional column vector of parameters. In this set up, the non-negative function φ_i is modeled via logit link function. This function is linear and other appropriate link functions that allow φ_i being negative may be used. In addition, in this paper we suppose that $\boldsymbol{\theta}_i$ and φ_i are not related.

2.1 Zero-inflated Poisson Model

We consider a zero-inflated Poisson regression model in which the response variable $Y_i (i = 1, \ldots, n)$ has the distribution

$$Pr(Y_i = y_i) = \begin{cases} \varphi_i + (1-\varphi_i)exp(-\lambda_i), & y_i = 0, \\ (1-\varphi_i)\exp(-\lambda_i)\lambda_i^{y_i}/y_i!, & y_i > 0, \end{cases} \tag{2}$$

where the parameter $\lambda_i(x_i)$ and φ_i satisfy $\log(\lambda_i) = \sum_{j=1}^{k} x_{ij}\beta_j$ and $0 < \varphi_i < 1$. The mean and the variance of the distribution are $E(\boldsymbol{Y}_i) = (1-\varphi_i)\lambda_i$ and $var(\boldsymbol{Y}_i) = (1-\varphi_i)\lambda_i(1 + \varphi_i\lambda_i)$.

2.2 Zero-inflated Model with Right Censoring

The value of response variable, Y_i, for some observations in a data set, may be censored. If censoring occurs for the ith observation, we have $Y_i \geq y_i$ (right censoring). However, if no censoring occurs, we know that $Y_i = y_i$. Thus, we can define an indicator variable d_i as

$$d_i = \begin{cases} 1 & \text{if } Y_i \geq y_i, \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

We can now write

$$Pr(Y_i \geq y_i) = \sum_{j=y_i}^{\infty} Pr(Y_i = j) = 1 - \sum_{j=0}^{y_i - 1} Pr(Y_i = j) \tag{4}$$

Therefore, the log-likelihood function of the censored zero-inflated regression model can be written as

$$\begin{aligned} \log L(\boldsymbol{\theta}_i; y_i) \quad &= \quad \sum_{i=1}^{n} \Big\{ (1 - d_i) \Big[I_{\{y_i=0\}} \log f(0; \boldsymbol{\theta}_i) + I_{\{y_i>0\}} \log f(y_i; \boldsymbol{\theta}_i) \Big] \\ & \qquad + d_i \log \Big(\sum_{j=y_i}^{\infty} Pr(Y_i = j) \Big) \Big\} \end{aligned} \tag{5}$$

We now calculate the log-likelihood function for the ZIPR model and by (2) and (5) we have

$$\begin{aligned} LL_{CZIP} \quad &= \quad \sum_{i=1}^{n} \Big\{ (1 - d_i) \Big[I_{y_i=0} \log\{\varphi_i + (1 - \varphi_i) \exp(-\lambda_i)\} \\ & \qquad + I_{y_i>0} \Big\{ \log(1 - \varphi_i) + y_i \log \lambda_i - \log(y_i!) - \lambda_i \Big\} \Big] \\ & \qquad + d_i \log \sum_{j=y_i}^{\infty} Pr(Y_i = j) \Big\} \end{aligned} \tag{6}$$

3 Parameter Estimation

In this section, we obtain the parameters estimation by the ML method. By taking the partial derivatives of (6) and setting them equal to zero, the likelihood equations for estimating β_r and δ_t are obtained. Thus we obtain

$$\begin{aligned} \frac{\partial LL_{CZIP}}{\partial \beta_r} \quad &= \quad \sum_{i=1}^{n} \Big\{ (1 - d_i) \Big[I_{\{y_i=0\}} \frac{-w_i^{-1} \exp(-\lambda_i)}{1 + w_i \exp(-\lambda_i)} x_{ir} \lambda_i + I_{\{y_i>0\}} (y_i - \lambda_i) x_{ir} \Big] \\ & \qquad + \frac{d_i}{\sum_{j=y_i}^{\infty} Pr(y_i = j)} \frac{\partial \sum_{j=y_i}^{\infty} Pr(y_i = j)}{\partial \beta_r} \Big\} = 0 \end{aligned} \tag{7}$$

$$\begin{aligned} \frac{\partial LL_{CZIP}}{\partial \delta_t} \quad &= \quad \sum_{i=1}^{n} \Big\{ (1 - d_i) \Big[I_{\{y_i=0\}} \frac{1 - \exp(\lambda_i)}{w_i + \exp(\lambda_i)} - I_{\{y_i>0\}} \Big] \frac{w_i}{1 + w_i} z_{it} \Big\} \\ &= \quad 0 \end{aligned} \tag{8}$$

where $w_i = \frac{\varphi_i}{1-\varphi_i} = \exp\{\sum_{j=1}^{m} z_{ij}\delta_j\}$. Furthermore, the expression for

$$\partial \sum_{j=y_i}^{\infty} Pr(y_i = j)/\partial\beta_r$$

is provided in the appendix.

It is clear that the likelihood equation (7) is nonlinear in the parameters. We can use an iterative technique to solve for the parameters in the above equations. The initial estimates of $\underline{\beta}$ and $\underline{\delta}$ may be taken as the corresponding final estimates of $\underline{\beta}$ and $\underline{\delta}$ from fitting a zero-inflated regression model to the data.

4 Goodness-of-fit Statistics

For ZI regression models, a measure of goodness of fit may be based on the deviance statistic D defined as

$$D = -2\big[\log L(\hat{\boldsymbol{\theta}}_i; \hat{\mu}_i) - \log L(\hat{\boldsymbol{\theta}}_i; y_i)\big] \tag{9}$$

where $\log L(\hat{\boldsymbol{\theta}}_i; \hat{\mu}_i)$ and $\log L(\hat{\boldsymbol{\theta}}_i; y_i)$ are the model's likelihood evaluated respectively under $\hat{\mu_i}$ and y_i. The log-likelihood function is available in equation (6).

For an adequate model, the asymptotic distribution of the deviance statistic D is chi-square distribution with $n-k-1$ degrees of freedom. Therefore, if the value for the deviance statistic D is close to the degrees of freedom, the model may be considered as adequate. When we have many regression models for a given data set, the regression model with the smallest value of the deviance statistic D is usually chosen as the best model for describing the given data.

In many data sets, the μ_i's may not be reasonably large and so the deviance statistic D may not be suitable. Thus, the log-likelihood statistic $\log(\hat{\boldsymbol{\theta}}_i; y_i)$ can be used as an alternative statistic to compare the different models. Models with the largest log-likelihood value can be chosen as the best model for describing the data under consideration.

5 Simulation Study

We conducted a simulation study in this section. All simulations were done using computer programs written in SAS codes. The parameter vector $(\beta_0, \beta_1, \beta_2, a_0, a_1)$ and the dispersion parameter were used in the simulation study. We fixed the parameter values as $\beta_0 = 0.1, \beta_1 = 0.5, \beta_2 = 0.5, a_0 = 0.1, a_1 = 1$ and we have considered a positive form for the dispersion parameter. For instance, based on CZIPR model, we have

$$\log(\lambda_i) = 0.1 + 0.5x_{1i} + 0.5x_{2i}, \quad logit(\varphi_i) = \log(\varphi_i[1-\varphi_i]^{-1}) = 0.1 + z_i, \tag{10}$$

where the variables x_{1i}, x_{2i} and z_i are generated from a continuous uniform $[0, 1]$, a continuous uniform $[0, 3.2]$ and a continuous uniform $[0, 0.1]$, respectively. Furthermore, we have chosen four censoring constants $C_1 = 4, C_2 = 6, C_3 = 8$ and $C_4 = 10$.

In this simulation study, we generated a set of data consisting of $n = 10000$ observations on three explanatory variables x_1, x_2 and z. In addition, the parameters $\beta_0, \beta_1, \beta_2, a_0$ and a_1 are estimated by the maximum likelihood method. As measures of goodness-of-fit, the

$-2LL$ ($-2\times$log-likelihood) and BIC are computed for this simulated data. Furthermore, we obtained the standard error for each parameter. Each standard error is reported in parentheses under its corresponding parameter estimation. Also, the percentages of censored y-values were computed for this simulation.

In all cases, censoring gives a better fit than the full model (uncensored model). The fit becomes better as the percentage of censoring increases.

6 Example

In this example, we fit the CZIPR model and ZIPR model to a count data set. This count data set is gathered by the state wildlife biologists and they would like to analyze how many fish are being caught by fishermen at a state park. The biologists asked from visitors how long they stayed, how many people were in the group, were there children in the group and how many fish were caught. There are excess zeros in the data because some visitors who did fish did not catch any fish.

Table 1: Parameter Estimation

	Parameter estimation and se					Goodness.of.fit		
model	a_0	a_1	b_0	b_1	b_2	$-2LL$	BIC	censored %
CZIPR	0.3446	0.4822	0.49670	0.71830	0.6232	20408	20454	17.59
	(0.0287)	(0.0323)	(0.0111)	(0.0494)	(0.8488)			
CZIPR	0.3136	0.4746	0.4935	0.3624	1.0347	25697	25743	8.04
	(0.0264)	(0.0283)	(0.0097)	(0.0449)	(0.7717)			
CZIPR	0.3006	0.4749	0.4936	0.2128	1.3078	28173	28219	3.24
	(0.0256)	(0.0270)	(0.0093)	(0.0434)	(0.7463)			
CZIPR	0.2972	0.474	0.4934	0.1539	1.4047	29195	29241	1.15
	(0.0253)	(0.0265)	(0.0092)	(0.0428)	(0.7369)			
Full	0.2995	0.4683	0.4924	0.1236	1.4413	29694	29740	-
	(0.0251)	(0.0263)	(0.0091)	(0.0425)	(0.7321)			

Table 2: Descriptive Statistics of the Variables

Variable	Mean	Std Dev	Min	Max	Variance
count	3.296	11.635	0	149	135.374
child	0.684	0.850	0	3	0.723
persons	2.528	1.113	1	4	1.238

The number of observation is 250 groups that went to a park. The response variable is how many fish each group caught (*count*) and the independent variables are how many children were in the group (*child*), how many people were in the group (*persons*), and whether or not they brought a camper to the park (*camper*). In Table 2, we can see the descriptive statistics of *count*, *child* and *persons* variables. Also, the frequency of variable *camper* is available in Table 3. Furthermore, Figure 1 is the histogram of the response variable.

Table 3: The Frequency of Variable Camper

camper	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	103	41.2	103	41.2
1	147	58.8	250	100.0

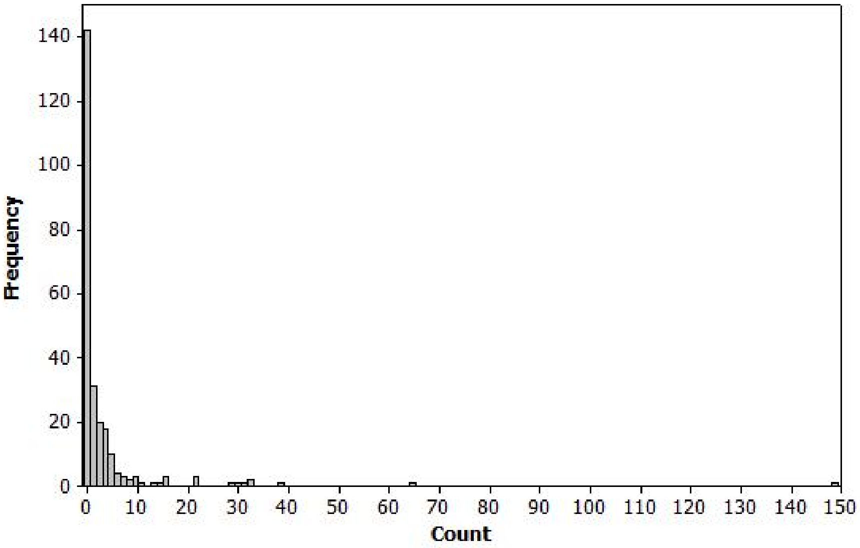

Figure 1: Histogram of the response variable

In this example, the dependent variable *count* has 108 zeros (43.2%). Therefore, the zero-inflated Poisson regression model will be adequate for analyzing this data set. However, the purpose of this example is to demonstrate censoring on the dependent variable *count*.

7 Discussion

We have used ZIPR model to analyze the complete data set without any censoring. Also, we have chosen five points as the censoring points (C) to see the effects of censoring on the parameter estimation, standard error and goodness-of-fit and we have used CZIPR model to analyze the censored data set. When the values of y_i are greater than or equal to C, we have censoring. Furthermore, We have computed the percentages of censored y-values according to censored points $[C = 4, 7, 10, 13 \ and \ 16]$ and the percentages are $18\%, 10\%, 7.2\%, 6.4\% \ and \ 4.8\%$, respectively.

The model is as follow,

$$\begin{aligned} \lambda &= \exp(b_0 + b_1 * camper + b_2 * persons + b_3 * child) \\ logit(\varphi) &= \log(\varphi[1-\varphi]^{-1}) = a_0 + a_1 * child \end{aligned} \quad (11)$$

Table 4: Parameter Estimation

	Parameter estimation and se						Goodness.of.fit		
model	a_0	a_1	b_0	b_1	b_2	b_3	$-2LL$	BIC	censored %
CZIPR	-0.2557	1.0930	-0.4606	0.47420	0.49070	-0.55910	471.3	504.5	18.0
	(0.2672)	(0.2849)	(0.2530)	(0.15630)	(0.07400)	(0.15210)			
CZIPR	-0.5400	1.1298	-0.4128	0.46290	0.51770	-0.57450	626.8	659.9	10.0
	(0.2568)	(0.2619)	(0.2249)	(0.13170)	(0.06163)	(0.12470)			
CZIPR	-0.5910	1.1326	-0.3775	0.45470	0.54990	-0.63640	718.6	751.7	7.2
	(0.2459)	(0.2538)	(0.2109)	(0.12140)	(0.05720)	(0.11580)			
CZIPR	-0.5836	1.1450	-0.3594	0.46390	0.57380	-0.65910	794.2	827.3	6.4
	(0.2368)	(0.2486)	(0.2013)	(0.11490)	(0.05415)	(0.10870)			
CZIPR	-0.6075	1.1503	-0.3362	0.47070	0.58910	-0.68530	872.5	905.6	4.8
	(0.2307)	(0.2443)	(0.1958)	(0.11070)	(0.05246)	(0.10560)			
Full	-0.9150	1.1857	-1.0572	0.77090	0.88860	-1.16750	1271.6	1304.7	-
	(0.2503)	(0.2654)	(0.1812)	(0.09384)	(0.04663)	(0.09471)			

From Table 4, when the censored percentages increase, the standard errors also increase for the censored models. Also, when we compare the b_i's, the censored model with less censoring percentage is the best in terms of standard error, however the full model has smaller standard error. Furthermore, according to a_i's, the censored model with the smallest censoring percentage is the best in terms of standard error, in addition, the standard errors for a_0 and a_1 for this model are also smaller than the full model.

When we compare the CZIPR and full model in Table 4, we can see that the goodness-of-fit statistics ($-2LL$ and BIC) for censored models increase as the censoring percentages decrease and also the $-2LL$ and BIC for all censored models are smaller than the full model.

The censored model CZIPR performed well in comparisons to the full model. The goodness-of-fit statistics ($-2LL$ and BIC) for censored model increase as the censoring percentages decrease. This implies that the fit is better when the percentages of censoring increase.

Appendix

From (4), $\sum_{j=y_i}^{\infty} Pr(Y_i = j) = 1 - \sum_{j=0}^{y_i-1} Pr(Y_i = j)$. So we have

$$\frac{\partial \sum_{j=y_i}^{\infty} Pr(Y_i = j)}{\partial \beta_r} = -\sum_{j=0}^{y_i-1} \frac{\partial Pr(Y_i = j)}{\partial \beta_r}$$

and

$$\frac{\partial \sum_{j=y_i}^{\infty} Pr(Y_i = j)}{\partial \alpha} = -\sum_{j=0}^{y_i-1} \frac{\partial Pr(Y_i = j)}{\partial \alpha}.$$

For CZIPR, we have

$$\frac{\partial Pr(Y_i = j)}{\partial \beta_r} = \frac{\partial Pr(Y_i = j)}{\partial \lambda_i} \frac{\partial \lambda_i}{\partial \beta_r} = Pr(Y_i = j)(y_i - \lambda_i)x_{ir}.$$

References

[1] Cameron, A.C. and Trivedi, P.K. *Regression Analysis of Count Data*. Cambridge University Press, Cambridge, UK. 1998.

[2] Ridout, M., Demetrio, C.G.B. and Hinde, J. Models for count data with many zeros. *Invited paper presented at the Nineteenth International Biometric Conference*. Cape Town, South Africa: 179-190. 1998.

[3] Lambert, D. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*. 1992. 34: 1-14.

[4] Caudill, S.B. and Mixon Jr., F.G. Modeling household fertility decisions: estimation and testing censored regression models for count data. *Empirical Econom.* 1995. 20: 183 196.

[5] Terza, J.V. A tobit-type estimator for the censored Poisson regression model. *Econom. Lett.* 1985. 18: 361-365.

[6] Consul, P.C. and Famoye, F. Generalized Poisson regression model. *Comm. Statist. Theory Methods*. 1992. 21: 81-109.

[7] Famoye, F. and Singh, K.P. Zero-inflated generalized Poisson model with an application to domestic violence data. *Journal of Data Science*. 2006. 4(1): 117-130.

[8] Famoye, F. and Wang, W. Censored generalized Poisson regression model. *Computational Statistics and Data Analysis*. 2004. 46: 547-560.

[9] Hall, D.B. Zero-inflated Poisson and binomial regression with random effects: a case study. *Biometrics*. 2000. 56: 1030-1039.