Sains Malaysiana 41(11)(2012): 1483–1487

Parameter Estimation on Zero-Inflated Negative Binomial Regression with Right Truncated Data

(Anggaran Parameter untuk Regresi Binomial Negatif Sifar-Melambung dengan Pemangkasan Data Sebelah Kanan)

SEYED EHSAN SAFFARI* & ROBIAH ADNAN

ABSTRACT

A Poisson model typically is assumed for count data, but when there are so many zeroes in the response variable, because of overdispersion, a negative binomial regression is suggested as a count regression instead of Poisson regression. In this paper, a zero-inflated negative binomial regression model with right truncation count data was developed. In this model, we considered a response variable and one or more than one explanatory variables. The estimation of regression parameters using the maximum likelihood method was discussed and the goodness-of-fit for the regression model was examined. We studied the effects of truncation in terms of parameters estimation, their standard errors and the goodness-of-fit statistics via real data. The results showed a better fit by using a truncated zero-inflated negative binomial regression model when the response variable has many zeros and it was right truncated.

Keywords: Maximum likelihood; truncated data; zero-inflated negative binomial

ABSTRAK

Model Poisson biasanya diandaikan untuk data bilangan, tetapi apabila terdapat banyak nilai sifar bagi pemboleh ubah bersandar yang disebabkan oleh penyerakan lampau, regresi binomial negatif dicadangkan sebagai regresi bilangan. Dalam artikel ini, model regresi binomial negatif sifar-melambung, dengan pemangkasan data bilangan pada sebelah kanan dibangunkan. Dalam model ini, kami mempertimbangkan satu pemboleh ubah bersandar dan satu atau lebih pemboleh ubah tak bersandar. Anggaran bagi parameter regresi menggunakan kaedah kemungkinan maksimum dibincang dan ujian penyuaian untuk model regresi diperiksa. Kesan pemangkasan dari segi penganggaran parameter dan ralat piawai dikaji menggunakan data sebenar. Keputusan menunjukkan penyuaian adalah lebih baik apabila menggunakan model regresi binomial negatif sifar melambung dengar pemangkasan di sebelah kanan apabila pemboleh ubah respons mempunyai banyak sifar dan dipangkas di sebelah kanan.

Kata kunci: Binomial negatif sifar-melambung; data pangkasan; kemungkinan maksimum

INTRODUCTION

The two most popular models for count data are the Poisson model and the binomial model. The Poisson model is used for the unbounded counts and the negative binomial model is used when the counts are bounded. The Poisson distribution has only one parameter, whereas the negative binomial distribution has two parameters. Due to this property, the negative binomial regression model is more flexible than Poisson regression model. Moreover, the Poisson distribution should have the same mean and variance value and this is not what happens in the real or simulated data. Thus, the negative binomial regression model can be used instead of Poisson regression model when the data under consideration is overdispersed.

However, often the counts show an excess of zeroes compared with what is expected. There are some researchers, who analyzed the zero-inflated count regression models. For instance, Lambert (1992) derived the zero-inflated Poisson regression model (ZIPR) model and its asymptotic properties of the ML estimator. Hall (2000) proposed the zero-inflated binomial (ZIB) regression model and studied random effects into ZIP and ZIB models. A zero-inflated generalized Poisson (ZIGP) regression model has been proposed by Famoye and Singh (2006).

There are many researches in which truncation problem is discussed. Most of these researches mentioned a zero-truncated count regression model such as zero-truncated Poisson or zero-truncated negative binomial regression model. Unlike these models, there are some situations in which there are so many zeros (zero inflation) in the data set. Because of the lack of information in a right truncated zero-inflated count regression model, in this paper we discuss the effect of a right truncation problem in a zero-inflated negative binomial regression model. The main objective of this paper was to introduce a right truncated zero-inflated negative binomial regression model to handle the zero-inflation and truncation problems together. The zero-inflated model is formulated in the next section that

follows, the zero-inflated negative binomial regression model is defined and the link functions are described. Next, the right truncated zero-inflated negative binomial model is discussed and the likelihood function is obtained. In the section that follows, the parameter estimation of the model is defined using maximum likelihood method. In the last section, the goodness-of-fit for the regression model is examined and a test statistic for examining.

THE MODEL

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

$$P\left(Y_i = y_i \middle| x_i, z_i\right) = \begin{cases} \varphi_i + \left(1-\varphi_i\right) f\left(0;\theta_i\right), & y_i = 0 \\ \left(1-\varphi_i\right) f\left(y_i;\theta_i\right) & y_i > 0 \end{cases}, \tag{1}$$

where $f(y_i; \theta_i)$, $y_i = 0,1,2,\ldots$ is the pdf of Y_i and $0 < \varphi_i < 1$. Furthermore, the function $\varphi_i = \varphi_i(z_i)$ satisfies $\text{logit}(\varphi_i) = \log(\phi_i[1-\varphi_i]^{-1}) = \sum_{j=1}^{m} z_{ij}\delta_j$ where $z_i = (z_{i1}, z_{i2}, \ldots, z_{im})$ the i-th row of covariate matrix Z and $\delta = (\delta_1, \delta_2, \ldots, \delta_m)$ 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 assumed that θ_i and φ_i are not related.

ZERO-INFLATED NEGATIVE BINOMIAL MODEL

We consider a zero-inflated negative binomial regression model in which the response variable $Y_i (i = \mathbf{1}, \ldots, n)$ has the distribution:

$$\begin{cases} \varphi_i + \left(1-\varphi_i\right)\left(\dfrac{\alpha^{-1}}{\alpha^{-1}+\mu_i}\right)^{\alpha^{-1}} & , y_i = 0 \\ \left(1-\varphi_i\right)\dfrac{\Gamma\left(y_i+\alpha^{-1}\right)}{\Gamma\left(y_i+1\right)\Gamma\left(\alpha^{-1}\right)}\left(\dfrac{\alpha^{-1}}{\alpha^{-1}+\mu_i}\right)^{\alpha^{-1}}\left(\dfrac{\mu_i}{\alpha^{-1}+\mu_i}\right)^{y_i} & , y_i > 0, \end{cases} \tag{2}$$

where $\alpha(\geq 0)$ is a dispersion parameter that is assumed not to depend on covariates. Furthermore, the model in (2) reduces to the ZIP distribution when the parameter $\alpha \to 0$ and the parameter $\lambda_i(X_i)$ and φ_i satisfy $\mathbf{log}(\lambda_i) = \sum_{j=1}^{m} x_{ij}\beta_j$ and $\mathbf{0} < \varphi i < \mathbf{1}$. The mean and the variance of the distribution are $E(Y_i) = (\mathbf{1} - \varphi_i)\mu_i$ and $\text{Var}(Y_i) = (\mathbf{1} - \varphi_i)\mu_i(\mathbf{1} + \varphi_i\mu_i + \alpha\mu_i)$.

ZERO-INFLATED MODEL WITH RIGHT TRUNCATION

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

$$f_T\left(y_i;\theta_i\right) = \frac{f\left(y_i;\theta_i\right)}{1-\Pr\left(Y_i \geq y_i\right)}, \quad i = 1,\ldots,k, \tag{3}$$

where k is the number of observation after truncation.

According to (3), we can write the log-likelihood function of the right truncated count regression model as follow:

$$\log L(\theta_i; y_i) = \sum_{i=1}^{k}\left[\log f\left(y_i;\theta_i\right) - \log\left(1-\Pr\left(Y_i \geq y_i\right)\right)\right]. \tag{4}$$

By taking partial derivatives respect to θ and equal to zero, we can obtain the parameter estimation. Furthermore, by replacing $f(y_i; \theta_i)$ into the negative binomial distribution, its distributions with right truncation will be obtained as follow:

$$\Pr\left(Y_i = y_i\right) = \begin{cases} \dfrac{\varphi_i + \left(1-\varphi_i\right)\left(\dfrac{\alpha^{-1}}{\alpha^{-1}+\mu_i}\right)^{\alpha^{-1}}}{1-\sum_{y_i=t_i+1}^{\infty}\left(1-\varphi_i\right)g\left(y_i;\mu_i,\alpha\right)}, & y_1 = 0 \\ \dfrac{\left(1-\varphi_i\right)g\left(y_i;\mu_i,\alpha\right)}{1-\sum_{y_i=t_i+1}^{\infty}\left(1-\varphi_i\right)g\left(y_i;\mu_i,\alpha\right)}, & 1 \leq y_i \leq t_u \end{cases}, \tag{5}$$

where,

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

and t_i is the truncation point for y_i which means that when $Y_i > t_i$ we truncate the response variable.

We can obtain the log-likelihood function for ZINBR model with right truncation as follow:

$$\begin{aligned} LL_{\{TZINB\}} = \sum_{i=1}^{k}\Bigg\{ & I_{\{y_i=0\}}\log\left(\varphi_i + \left(1-\varphi_i\right)\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu_i}\right)^{\alpha^{-1}}\right) \\ & -\log\left(1-\sum_{y_i=t_i+1}^{\infty}\left(1-\phi_i\right)g\right) \\ & +I_{\{1<y_i\leq t_i\}}\left[\log\left(1-\varphi_i\right)+\log g-\log\left(1-\sum_{y_i=t_i+1}^{\infty}\left(1-\varphi_i\right)g\right)\right]\Bigg\} \end{aligned} \tag{6}$$

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

$$\begin{aligned} \log g\left(y_i;\mu_i,\alpha\right) = & \sum_{j=0}^{y_i-1}\log\left(j+\alpha^{-1}\right) - \log y_i\,! + \\ & y_i\log\alpha\mu_i - y_i\log\left(1+\alpha\mu_i\right) \end{aligned}$$

PARAMETER ESTIMATION

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

$$\frac{\partial LL_{(TZINB)}}{\partial \beta_r} = \sum_{i=1}^{k}\left\{ I_{\{y_i=0\}}\left[\frac{-\left(1+\alpha\mu_i\right)^{-\alpha^{-1}-1}}{w_i+\left(1+\alpha\mu_i\right)^{-\alpha^{-1}}}+ \frac{\sum_{y=t_i+1}^{\infty}\left(1-\varphi_i\right)g\,\frac{y_i+2\alpha\mu_i y_i-\mu_i}{\mu_i\left(1+\alpha\mu_i\right)}}{1-\sum_{y_i=t_i+1}^{\infty}\left(1-\varphi_i\right)g}\right]x_{ir}\mu_i \right.$$
$$+ I_{\{1\le y_i\le t_i\}}\left[y_i\left(\frac{1}{\mu_i}-\frac{\alpha}{1+\alpha\mu_i}\right)+ \left.+\frac{\sum_{y_i=t_i+1}^{\infty}\left(1-\varphi_i\right)g\,\frac{y_i+2\alpha\mu_i y_i-\mu_i}{\mu_i\left(1+\alpha\mu_i\right)}}{1-\sum_{y_i=t_i+1}^{\infty}\left(1-\varphi_i\right)g}\right]x_{ir}\mu_i\right\}=0 \tag{7}$$

$$\frac{\partial LL_{(TZINB)}}{\partial \alpha} = \sum_{i=1}^{k}\left\{ I_{\{y_i=0\}}\left[\frac{\alpha^{\{-1\}\log\left(1+\alpha\mu_i\right)}-\frac{\mu_i}{1+\alpha\mu_i}}{w_i+\left(1+\alpha\mu_i\right)^{-\alpha^{-1}}}\frac{1}{\alpha}\right.\right.$$
$$+\left(\sum_{y_i=t_i+1}^{\infty}\left(1-\varphi_i\right)\right)\left(\frac{\Gamma'\left(y_i+\alpha^{-1}\right)}{\Gamma\left(y_i+\alpha^{-1}\right)}-\frac{\Gamma'\left(\alpha^{-1}\right)}{\Gamma\left(\alpha^{-1}\right)}\right.$$
$$+\alpha^{-2}\log\left(1+\alpha\mu_i\right)$$
$$\left.-\frac{\alpha^{-1}}{1+\alpha\mu_i}\left(\mu_i-y_i\right)\right)g\Bigg)\Bigg/\left(1-\sum_{y_i=t_i+1}^{\infty}\left(1-\varphi_i\right)g\right)$$
$$+I_{\{1\le y_i\le t_i\}}\left[-\sum_{j=0}^{y_i-1}\frac{\alpha^{-2}}{j+\alpha^{-1}}-\frac{y_i\mu_i}{1+\alpha\mu_i}+\alpha^{\{-1\}y}\right]$$
$$+\left(\sum_{y_i=t_i+1}^{\infty}\left(1-\varphi_i\right)\right)\left(\frac{\Gamma'\left(y_i+\alpha^{-1}\right)}{\left(y_i+\alpha^{-1}\right)}-\frac{\Gamma'\left(\alpha^{-1}\right)}{\Gamma\left(\alpha^{-1}\right)}\right.$$
$$+\alpha^{-2}\log\left(1+\alpha\mu_i\right)$$
$$\left.\left.-\frac{\alpha^{-1}}{1+\alpha\mu_i}\left(\mu_i-y_i\right)\right)g\Bigg)\Bigg/\left(1-\sum_{y_i=t_i+1}^{\infty}\left(1-\varphi_i\right)g\right)\right\}$$
$$=0 \tag{8}$$

$$\frac{\partial LL_{(TZINB)}}{\partial \delta_s} = \sum_{i=1}^{k}\left\{ I_{\{y_i=0\}}\left[\frac{1-\left(1+\alpha\mu_i\right)^{-\alpha^{-1}}}{w_i+\left(1+\alpha\mu_i\right)^{-\alpha^{-1}}}\right.\right.$$
$$\left.-\frac{\sum_{y_i=t_i+1}^{\infty}g}{\left(1-\varphi_i\right)^{\{-1\}}-\sum_{y_i=t_i+1}^{\infty}g}\right]\varphi_i z_{is}$$
$$\left.-I_{\{1\le y_i\le t_i\}}\left[1+\frac{\sum_{y_i=t_i+1}^{\infty}g}{\left(1-\varphi_i\right)^{\{-1\}}-\sum_{y_i=t_i+1}^{\infty}g}\right]\varphi_i z_{is}\right\}=0 \tag{9}$$

GOODNESS-OF-FIT STATISTICS

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

$$D = -2[\log L\left(\hat{\theta}_l;\hat{\mu}_l\right) - \log L\left(\hat{\theta}_l;y_i\right)] \tag{10}$$

where logs $L\left(\hat{\theta}_l;\hat{\mu}_l\right)$ and log $L\left(\hat{\theta}_l;y_i\right)$ are the model's likelihood evaluated respectively under $\hat{\mu}_l$ 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 $\hat{\mu}_l$'s may not be reasonably large and so the deviance statistic D may not be suitable. Thus, the log-likelihood statistic log $L(\hat{\theta}_l;y_i)$ 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.

AN APPLICATION

The state wildlife biologists want to model how many fish are being caught by fishermen at a state park. Visitors are asked how long they stayed, how many people were in the group, were there children in the group and how many fish were caught. Some visitors do not fish, but there is no data on whether a person fished or not. Some visitors who did fish did not catch any fish so there are excess zeros in the data because of the people that did not fish. We have data on 250 groups that went to a park. Each group was questioned about how many fish they caught (count), 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 addition, to predict the number of fish caught, there is interest in predicting the existence of excess zeroes, i.e. the zeroes that were not simply a result of bad luck fishing. We will use the variables child, persons and camper in our model. Table 1 shows the descriptive statistics of using variables and also the camper variable which has two values, zero and one as Table 2.

Figure 1 shows the histogram of the *count* variables and it is clear that we have zero-inflation problem, also we have few large values that we are interested to truncate them. We have considered the model as follow:

$$\log\lambda = b_0 + b_1 * camper + b_2 * persons + b_3 * child,$$
$$logit\ \phi = a_0 + a_1 * child$$

TABLE 1. Descriptive statistics

Variable	Mean	Std Dev	Min	Max	Variance
Count	3.296	11.6350281	0	149	135.3738795
Child	0.684	0.8503153	0	3	0.7230361
Persons	2.528	1.1127303	1	4	1.2381687

TABLE 2. Camper variable

Camper	Frequency	Percent
0	103	41.2
1	147	58.8

Furthermore, we put three truncation points, t_1 = 3, t_2 = 5, t_3 = 10. Table 3 shows the estimation of the parameters according to different truncation constants. Also, the $-2LL$ and *AIC* are presented as the goodness-of-fit measures. According to the truncation points, there is 22.8% truncated data when t_1 = 3, 12% when t_2 = 5 and 7.2% when t_3 = 10.

CONCLUSION

In this article we showed that the zero-inflated negative binomial regression model can be used to fit right truncated data. In Table 1, the percentage of zeros of the response variable is 56.8%. This kind of data is defined as zero-inflated data. The zero-inflated negative binomial regression model with right truncation (TZINBR) is fitted to these real data. The results from the fish data are summarized in Tables 1 to 3. The goodness-of-fit measures are presented in the Table 3 according to different truncation points and it is obvious that we have a smaller value for $-2LL$ or *AIC* when the percentage of truncation increase and that is because of the number of the data which are used in the model.

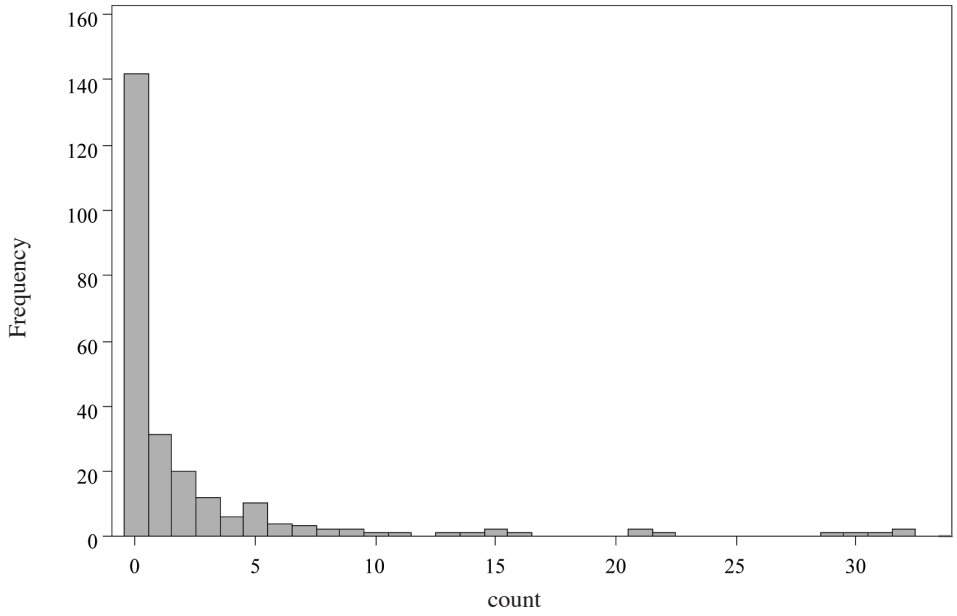

FIGURE 1. Histogram of *count* variable

TABLE 3. Parameter estimation

	Parameter estimation							Goodness-of-fit	
	b_0	b_1	b_2	b_3	a_0	a_1	α	$-2LL$	*AIC*
$t_1 = 3$	-2.0093 (0.7619)	0.4316 (0.4801)	0.9553 (0.4834)	-1.0352 (0.6861)	-2.9554 (1.5158)	2.3820 (1.0541)	0.5923 (0.6484)	258.7	272.7
$t_2 = 5$	-1.6136 (0.5123)	0.9865 (0.4577)	0.8010 (0.2919)	-1.5963 (0.7940)	-5.9236 (7.6520)	3.2780 (3.4275)	1.4435 (0.6594)	393.9	407.9
$t_3 = 10$	-1.7913 (0.4147)	0.8429 (0.3207)	0.9345 (0.1857)	-1.3528 (0.4458)	-4.2706 (2.1359)	2.6254 (1.0505)	1.1578 (0.3872)	560.4	574.4

REFERENCES

Cameron, A.C. & Trivedi, P.K. 1998. *Regression Analysis of Count Data*. Cambridge, UK: Cambridge University Press.

Famoye, F. & Singh, K.P. 2006. Zero-inflated generalized Poisson model with an application to domestic violence data. *Journal of Data Science* 4(1): 117-30.

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

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

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

Saffari, S.E. & Robiah Adnan 2010. Zero-Inflated Negative Binomial Regression Model with Right Censoring Count Data. *Proceedings of the Faculty of Science Postgraduate Conference (FSPGC'10)*; October 5-7, Johor, Malaysia.

Department of Mathematical Sciences
Faculty of Science
Universiti Teknologi Malaysia
81310 Skudai, Johor
Malaysia

*Corresponding author; email: ehsanreiki@yahoo.com

Received: 25 May 2012
Accepted: 2 July 2012