Zero-Inflated Negative Binomial Regression Model with Right Censoring Count Data

Seyed Ehsan Saffari and Robiah Adnan
Department of Mathematics, Faculty of Science, Universiti Teknologi Malaysia, Malaysia

Received: January 16, 2011 / Accepted: March 01, 2011 / Published: September 25, 2011.

Abstract: A Poisson model is typically assumed for count data, but when there are so many zeros 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 censoring count data is developed. 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 censoring in terms of parameters estimation and their standard errors via simulation.

Key words: Zero-inflated negative binomial, censored data, maximum likelihood, simulation.

1. Introduction

The two most popular models for count data are the Poisson model and the binomial model. The Poisson model is sued 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. Further, 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 to what is expected. There are some researches to be done to analyze the zero-inflated count regression models. For instance, Lambert [1] derived the zero-inflated Poisson regression (ZIPR) model and its asymptotic properties of the ML estimator. Hall [2] 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 [3].

In many applications, the response variable of the count data is often censored. Famoye & Wang [4] suggested the use of censored generalized Poisson regression (CGPR) model. They applied the (CGPR) model to a data set on fertility from the Michigan Panel Study of Income Dynamics and compared it to the Poisson, truncated Poisson, (GPR), and censored Poisson regression models. The censored generalized Poisson regression (CGPR) model illustrated how estimates of parameters can be greatly improved if censoring is considered in data set that is overdispersed or underdispersed. Cameron & Trivedi [5] suggested models allowing for censoring count data.

In this article, the main objective is to explain how we can use zero-inflated negative binomial regression model in right censored data. In section 2, the zero-inflated negative binomial regression model is defined and the likelihood function of zero inflated regression model in right censored data is formulated.

Corresponding author: Seyed Ehsan Saffari, Ph.D. student, research field: statistics. E-mail: Esseyed3@live.utm.my.
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 negative binomial regression model in terms of the parameter estimation, standard errors and goodness-of-fit statistic is conducted in section 5.

2. 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

$$P(Y_i = y_i | x_i, \theta) = \begin{cases} \phi_i + (1 - \phi_i) f(y_i; \theta_i), & y_i = 0, \\ (1 - \phi_i) f(y_i; \theta_i), & y_i > 0, \end{cases}$$

Where $f(y_i; \theta_i), y_i = 0, 1, 2, \ldots$ is the pdf of $Y_i$ and $0 < \phi_i < 1$. Furthermore, the function $\phi_i = \phi_i(z_i)$ satisfies $logit(\phi_i) = logit(\phi_i[1 - \phi_i])^{-1} = \sum_{j=1}^{m} z_{ij} \delta_j$ where $z_i = (z_{i1}, z_{i2}, \ldots, z_{im})$ is 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 setup, the non-negative function $\varphi_i$ is modeled via logit link function. This function is linear and other appropriate link functions that allow $\varphi_i$ being negative may be used. In addition, in this paper we suppose that $\theta_i$ and $\varphi_i$ 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

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

where the parameter $\lambda_i$ and $\varphi_i$ 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 $E(Y_i) = (1 - \varphi_i) \lambda_i$ and $Var(Y_i) = (1 - \varphi_i) \lambda_i(1 + \varphi_i \lambda_i)$.

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 i-th 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

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

We can now write

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

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

$$log L(\theta_i ; y_i) = \sum_{i=1}^{n} \left[ (1 - d_i) [I_{y_i = 0} log(f(0; \theta_i)) + I_{y_i > 0} log(f(y_i; \theta_i))] + d_i \log \left( \sum_{j=0}^{\infty} Pr(Y_i = j) \right) \right]$$

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

$$LL_{ZIPR} = \sum_{i=1}^{n} \left[ (1 - d_i) \left[ I_{y_i = 0} log(\varphi_i + (1 - \varphi_i) \left(\frac{\alpha - 1}{\alpha - \gamma} + \gamma \right)^{-1}) \right] + I_{y_i > 0} \left[ \log(1 - \varphi_i) + \sum_{j=1}^{y_i-1} \log(j + \alpha - 1) - \log y_i! + y_i \log \alpha \mu_i - y_i \log(1 + \alpha \mu_i) \right] + d_i \log \sum_{j=0}^{\infty} Pr(Y_i = j) \right]$$

3. Parameter Estimation

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

$$\frac{\partial LL_{ZIPR}}{\partial \beta_i} = \sum_{i=1}^{n} \left[ (1 - d_i) \left[ I_{y_i = 0} \left(\frac{\alpha - 1}{\alpha - \gamma} + \gamma \right)^{-1} \left( \frac{1}{\mu_i} - \frac{\alpha}{1 + \alpha \mu_i} \right) \right] x_{i+ \mu_i} + I_{y_i > 0} \left( \frac{d_i}{\sum_{j=0}^{\infty} Pr(y_i = j) \partial \beta_i} \right) \right] = 0 \quad (7)$$

$$\frac{\partial LL_{ZIPR}}{\partial \delta_i} = \sum_{i=1}^{n} \left[ (1 - d_i) \left[ I_{y_i = 0} \left(\frac{\alpha - 1}{\alpha - \gamma} + \gamma \right)^{-1} - I_{y_i > 0} \right] \frac{y_i}{\mu_i} \right] = 0 \quad (8)$$
where
\[
\frac{\partial L_{\text{ZINB}}}{\partial \theta} = \sum_{i=1}^{n} \left\{ (1 - d_i) \left[ I_{y_i=1} \log \left( \frac{1 + \alpha \theta_i}{1 + \alpha \theta_i} \right) - \frac{\mu_i}{1 + \alpha \theta_i} \right] + I_{y_i>0} \left\{ - \frac{\gamma y_i}{\alpha + \gamma} - \frac{y_i \theta_i}{1 + \alpha \theta_i} + \frac{y_i}{\alpha} \right\} \right\} = 0
\]
where \( w_i = \frac{x_i}{1 + \alpha \theta_i} = \exp \left\{ \sum_{j=1}^{m} \frac{y_i}{\alpha} \right\} \). Furthermore, the expression for \( \frac{\partial \sum_{j=y_i} \Pr(y_i = j)}{\partial \alpha} \) is provided in the appendix.

It is clear that the likelihood Eq. (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 \( \beta \) and \( \delta \) may be taken as the corresponding final estimates of \( \beta \) and \( \delta \) 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
\[
D = -2 \left[ \log L(\hat{\beta}; \mu_i) - \log L(\hat{\beta}; y_i) \right]
\]

where \( \log L(\hat{\beta}; \mu_i) \) and \( \log L(\hat{\beta}; y_i) \) are the model’s likelihood evaluated respectively under \( \mu_i \) and \( y_i \). The log-likelihood function is available in Eq. (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 \( \mu_i \)’s may not be reasonably large and so the deviance statistic \( D \) may not be suitable. Thus, the log-likelihood statistic \( \log L(\hat{\beta}; y_i) \) 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 based on samples obtained using bootstrap method. 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 \( \beta_0 = \beta_1 = \beta_2 = 1, a_0 = 0.5, a_1 = 1.5 \) and we consider three forms (positive, zero and negative) for the dispersion parameter. For instance, based on CZINBR model, we have
\[
\log(\lambda_i) = 1 + x_{1i} + x_{2i},
\]
\[
\logit(\varphi_i) = \log(\varphi_i | 1 - \varphi_i |^{-1}) = 0.5 + 1.5z_i
\]
where the variables \( x_{1i}, x_{2i} \) and \( z_i \) are generated from a continuous uniform \([0, 1]\), a continuous uniform \([1, 2]\) and a continuous uniform \([0, 1]\), respectively. For all simulations, we chose a censoring constant \( C = 30 \).

In this simulation study, we generated a set of data consisting of \( n = 200 \) observations on three explanatory variables \( x_1, x_2 \) and \( z \) with replication10. In each simulation, 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 \( -2LL, (-2 \times \text{log-likelihood}) \) is computed for each 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 zero-inflated and censored \( y \)-values were computed for each simulation. The results are summarized in Table 1.

| Parameter Estimation of CZINBR. | Goodness-of-fit |
|---------------------------------|-----------------|
| \( \beta_0 \) | \( \beta_1 \) | \( \beta_2 \) | \( a_0 \) | \( a_1 \) | Dispersion | \(-2LL\) | ZI% | Censored% |
| Positive | Zero | Negative |
| \( \beta_0 \) | 2.0066 | 0.6032 | 0.5056 | 1.0780 | 1.4799 | Positive | 3.150.0 | 75.65 | 4.5 |
| \( \beta_1 \) | (0.0832) | (0.0456) | (0.0515) | (0.1225) | (0.2488) | Zero | 2.951.9 | 76.85 | 4.3 |
| \( \beta_2 \) | 1.6057 | 0.7865 | 0.7759 | 0.9067 | 1.6994 | Negative | 2.810.3 | 77.75 | 3.3 |
| \( a_0 \) | (0.0936) | (0.0577) | (0.0593) | (0.1278) | (0.2534) | | | | |
| \( a_1 \) | 1.5398 | 0.9394 | 0.8461 | 1.0919 | 1.4382 | | | | |
| | (0.0958) | (0.0605) | (0.0579) | (0.1324) | (0.2433) | | | | |
6. Conclusions

In this article, we want to show that the zero-inflated negative binomial regression model can be used to fit right censored data. In Table 1, the percentages of zeros are approximately between 75% to 80%. This kind of data is defined as zero-inflated data. The zero-inflated negative binomial regression model with right censoring (CZINBR) is fitted to these simulated data. The results from the simulation are summarized in Table 1. The fit is best when the dispersion is negative irrespective of the sample size. The fit is worst when the dispersion is positive.

References

[1] D. Lambert, Zero-inflated Poisson regression, with an application to defects in manufacturing, Technometrics 34 (1992) 1-14.
[2] D.B. Hall, Zero-inflated Poisson and binomial regression with random effects: a case study, Biometrics 56 (2000) 1030-1039.
[3] F. Famoye, K.P. Singh, Zero-inflated generalized Poisson model with an application to domestic violence data, Journal of Data Science 4 (1) (2006) 117-130.
[4] F. Famoye, W. Wang, Censored generalized Poisson regression model, Computational Statistics and Data Analysis 46 (2004) 547-560.
[5] A.C. Cameron, P.K. Trivedi, Regression Analysis of Count Data, Cambridge University Press, Cambridge, UK, 1998.

Appendix

From Eq. (4), \( \sum_{j=0}^{\infty} Pr(Y_i = j) = 1 - \sum_{j=0}^{y_i-1} Pr(Y_i = j) \). So we have \( \frac{\partial \sum_{j=0}^{y_i} Pr(Y_i = j)}{\partial \alpha} = -\sum_{j=0}^{y_i-1} \frac{\partial Pr(Y_i = j)}{\partial \alpha} \) and \( \frac{\partial \sum_{j=0}^{y_i} Pr(Y_i = j)}{\partial \beta_r} = -\sum_{j=0}^{y_i-1} \frac{\partial Pr(Y_i = j)}{\partial \beta_r} \).

For CZINBR, we have

\[
\frac{\partial Pr(Y_i = j)}{\partial \beta_r} = \frac{\partial Pr(Y_i = j)}{\partial \mu_i} \frac{\partial \mu_i}{\partial \beta_r} = Pr(Y_i = j) \left[ \frac{\mu_i y_i}{(\alpha^{-1} + \mu_i)^2 - 1} \right] \frac{\alpha^{-1}}{\alpha^{-1} + \mu_i} \mu_i x_{ir},
\]

\[
\frac{\partial Pr(Y_i = j)}{\partial \alpha} = Pr(Y_i = j) \left[ \frac{-1}{\alpha^2} \log \left( \frac{\alpha^{-1}}{\alpha^{-1} + \mu_i} \right) + \frac{y_i - 1}{1 + \alpha \mu_i} - \frac{y_i}{\Gamma(y_i + \alpha^{-1})} + \frac{\Gamma(y_i + \alpha^{-1})}{\Gamma(y_i + \alpha^{-1})} \right].
\]