Jurnal Teknologi

Full paper

A Poisson Regression Model For Analysis of Censored Count Data with Excess Zeroes

Seyed Ehsan Saffari[a*], Robiah Adnan[a], William Greene[b], Maizah Hura Ahmad[a]

[a]*Department of Mathematical Sciences, Faculty of Science, Universiti Teknologi Malaysia, 81310 UTM Johor Bahru, Johor, Malaysia*
[b]*Department of Economics, Stern School of Business, New York University 44 West 4th St., New York, NY, 10012, USA*

*Corresponding author: esseyed3@live.utm.my

Article history

Received :21 January 2013
Received in revised form :
7 May 2013
Accepted :25 June 2013

Graphical abstract

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

Abstract

Typically, a Poisson regression model is assumed for count data. In many cases, there are many zeros in the dependent variable, therefore the mean is not equal to the variance value of the dependent variable. Thus, we suggest using a hurdle and zero-inflated Poisson regression model. Furthermore, the response variable in such cases is censored for some values. In this paper, a censored hurdle Poisson regression model and a censored zero-inflated Poisson regression model will be discussed to handle the overdispersion problem when there are excess zeros in the response variable. The estimation of regression parameters using the maximum likelihood method is discussed and the goodness-of-fit statistics for the regression model are examined. An example and a simulation will be used to compare the censored hurdle Poisson regression model with the censored zero-inflated Poisson regression model in terms of the parameter estimation, standard errors and the goodness-of-fit statistics.

Keywords: Censored model; Poisson regression; overdispersion; excess zeros

Abstrak

Biasanya, model regresi Poisson diandaikan untuk data kiraan. Dalam banyak kes, terdapat banyak sifar dalam pemboleh ubah bersandar, maka min tidak sama dengan nilai varians pemboleh ubah bersandar. Oleh itu, kami cadangkan penggunaan model "hurdle and zero-inflated Poisson regression." Tambahan pula, pemboleh ubah bersandar dalam kes-kes seperti itu ditapis untuk beberapa nilai. Dalam kertas kerja ini, model "censored hurdle Poisson regression" dan model "censored zero-inflated Poisson regression" akan dibincangkan untuk menangani masalah "overdispersion" apabila terdapat sifar yang berlebihan dalam pemboleh ubah bersandar. Anggaran parameter regresi menggunakan kaedah kebolehjadian maksimum dibincangkan dan statistik kebaikan kesesuaian bagi model regresi diperiksa. Satu contoh dan satu simulasi akan digunakan untuk membandingkan model "censored hurdle Poisson regression" dengan model "censored zero-inflated Poisson regression" dari segi anggaran parameter, ralat piawai dan statistik kebaikan kesesuaian.

Kata kunci: Censored model; regresi Poisson; *overdispersion*; sifar berlebihan

1.0 INTRODUCTION

Commonly, the starting point for the modeling of the number of reported claims is the Poisson distribution:

$$f_{Y_i}(y_i) = \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!} \qquad (1)$$

where Y_i is the response variable which follows a Poisson distribution with mean value of λ_i and the covariates are included in the model by the parameter $\lambda_i = \exp(\mathrm{x}_i'\beta)$. The Poisson distribution is equidispersed since its mean and variance are both equal to λ_i. In some cases, the number of zeros in the response variable is more than expected, thus the Poisson distribution cannot handle this kind of data anymore. In this case, there are two suggested models, hurdle model and zero-inflated model.

Mullahy (1986) has first discussed hurdle count data models. Hurdle models permit for a systematic difference in the statistical process governing individuals (observations) below the hurdle and individuals above the hurdle. In particular, a hurdle model is mixed by a binary outcome of the count being below or above the hurdle (the selection variable), with a truncated model for outcomes above the hurdle. That is why hurdle models sometimes are also called as two-part models.

Lambert (1992) introduced the original formulation for the zero-inflated Poisson (ZIP) model. She also talked about the models' extension from the Poisson and negative binomial and argued about the derivation of the maximum likelihood (ML) estimates. In her work, she discussed some simulations to test the sufficiency of the model. She also concluded that these

simulations with one covariate are suitable for both λ (parameter of the Poisson model) and p (the parameter of the logit part of the model).

The hurdle model is flexible and can handle both under- and overdispersion problem. A generalized hurdle model is introduced by Gurmu (1998) for the analysis of overdispersed or underdispersed count data. Greene (2007) has discussed about the comparison between hurdle and zero-inflated models as two part-models. He also discussed the hurdle and zero-inflated regression models with and without covariates.

In many applications, count data are often censored from above or below a specific point or a combination of them and this is because there are some outliers in the model which these outliers can be some large values or some small values. The case of variable threshold was considered by Caudill and Mixon (1995). To analyze censored data with a constant censoring threshold, Terza (1985) proposed the censored Poisson regression model and obtained the ML estimator using the Newton-Raphson method. In practice, censored count data are too dispersed to use the censored Poisson model.

In this article, we would like to study two problems together. The first problem is many zeros in the response variable that we suggested using a hurdle Poisson (HP) regression model or ZIP regression model. The second problem is the existing of some outliers or some large values in the response variable that we proposed to censor the data from the right side. In this paper, the main objective is to compare a censored hurdle Poisson (CHP) Regression model with a censored zero-inflated Poisson (CZIP) regression model. In section 2, the hurdle Poisson (HP) and ZIP regression models are defined and the likelihood function of the models 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 statistics for the regression models is examined. A simulation for the CHP and CZIP regression models in terms of the parameter estimation, standard errors and goodness-of-fit statistic is conducted in section 5.

■2.0 MATERIALS AND METHODS

2.1 HP Model

Let $Y_i, i = 1,2,\dots,n$ be a nonnegative integer-valued random variable and suppose $Y_i = 0$ is observed more than expected. We consider a hurdle Poisson regression model in which the response variable $Y_i (i = 1,\dots,n)$ has the distribution

$$P(Y_i = y_i) = \begin{cases} w_0 & y_i = 0 \\ (1 - w_0)\dfrac{e^{-\lambda_i}\lambda_i{}^{y_i}}{(1 - e^{-\lambda})y_i!} & y_i > 0 \end{cases} \quad (2)$$

where $0 < w_0 < 1$ and $w_0 = w_0(z_i)$ satisfy

$$\text{logit}(w_0) = \log\left(\frac{w_0}{1 - w_0}\right) = \sum_{j=1}^{m} z_{ij}\delta_j \quad (3)$$

where $z_i = (z_{i1} = 1, z_{i2}, \dots, z_{im})$ is the i-th row of covariate matrix Z and $\delta = (\delta_1, \delta_2, \dots, \delta_m)$ are unknown m-dimensional column vector of parameters. In this set up, the non-negative function w_0 is modeled via logit link function. This function is linear and other appropriate link functions that allow w_0 being negative may be used. In addition, there is interest in capturing any systematic variation in λ_i, the value of λ_i is most commonly placed within a loglinear model

$$\log(\lambda_i) = \sum_{j=1}^{m} x_{ij}\beta_j \quad (4)$$

and β_j's are the independent variables in the regression model and m is the number of these independent variables.

2.2 ZIP Model

Let Y_i be a nonnegative integer-valued random variable and suppose that there are many zeros in the response variable even more than what would typically be predicted. We consider a zero-inflated Poisson regression model in which the response variable $Y_i (i = 1, \dots, n)$ has the distribution

$$Pr(Y_i = y_i) = \begin{cases} \phi_i + (1 - \phi_i)\exp(-\lambda_i), & y_i = 0 \\ (1 - \phi_i)\dfrac{\exp(-\lambda_i)\lambda_i^{y_i}}{y_i!}, & y_i > 0 \end{cases} \quad (5)$$

where the parameter λ_i and ϕ_i satisfy $\log(\lambda_i) = \sum_{j=1}^{k} x_{ij}\beta_j$ and $0 < \phi_i < 1$.

2.3 Censored Model

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

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

We can now write

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

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

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

Thus, we can write the log-likelihood function for the censored HP and ZIP as follows,

$$LL_{CHP} = \sum_{i=1}^{n} \Bigg\{ (1 - d_i)\big[I_{y_i=0}\log w_0 + I_{y_i>0}\{\log(1 - w_0) - \lambda_i + y_i \log \lambda_i - \log(y_{i!}) - \log(1 - e^{-\lambda_i})\}\big] + d_i \log \sum_{j=y_i}^{\infty} Pr(Y_i = j)\Bigg\}$$

$$LL_{CZIP} = \sum_{i=1}^{n} \Bigg\{ (1 - d_i)\big[I_{y_i=0}\log\{\phi_i + (1 - \phi_i)\exp(-\lambda_i)\} + I_{y_i>0}\{\log(1 - \phi_i) - \lambda_i + y_i \log \lambda_i - \log(y_{i!})\}\big] + d_i \log \sum_{j=y_i}^{\infty} Pr(Y_i = j)\Bigg\}$$

2.4 Parameter Estimation

In this section, we obtain the parameters estimation by the ML method. By taking the partial derivative of the likelihood function and setting it equal to zero, the likelihood equation for estimating

the parameter is obtained. Thus we obtain the likelihood equations for the censored HP and ZIP regression models as follows,

$$\frac{\partial LL_{CHP}}{\partial \beta_r} = \sum_{i=1}^{n}\left\{(1-d_i)I_{y_i>0}\left[y_i - \lambda_i - \frac{\lambda_i e^{-\lambda_i}}{1-e^{-\lambda_i}}\right]x_{ir} + \frac{d_i}{\sum_{j=y_i}^{\infty} Pr(Y_i=j)}\frac{\partial \sum_{j=y_i}^{\infty} Pr(Y_i=j)}{\partial \beta_r}\right\} = 0$$

$$\frac{\partial LL_{CHP}}{\partial \delta_t} = \sum_{i=1}^{n}\{(1-d_i)[I_{y_i=0}(1-w_0) - I_{y_i>0}w_0]\}z_{it} = 0$$

where

$$\frac{\partial \sum_{j=y_i}^{\infty} Pr(Y_i=j)}{\partial \beta_r} = \sum_{j=y_i}^{\infty}(1-w_0)\frac{e^{-\lambda_i}{\lambda_i}^{j}}{(1-e^{-\lambda})^{-2}j!}\left[j\left(1-e^{-\lambda_i}\right) - \lambda_i\right]x_{ir}$$

and

$$\frac{\partial LL_{CZIP}}{\partial \beta_r} = \sum_{i=1}^{n}\left\{(1-d_i)\left[I_{y_i=0}\frac{-w_i^{-1}e^{-\lambda_i}}{1+w_i e^{-\lambda_i}}x_{ir}\lambda_i + I_{y_i>0}(y_i-\lambda_i)x_{ir}\right] + \frac{d_i}{\sum_{j=y_i}^{\infty} Pr(Y_i=j)}\frac{\partial \sum_{j=y_i}^{\infty} Pr(Y_i=j)}{\partial \beta_r}\right\} = 0$$

$$\frac{\partial LL_{CZIP}}{\partial \delta_t} = \sum_{i=1}^{n}\left\{(1-d_i)\left[I_{y_i=0}\frac{1-e^{\lambda_i}}{w_i+e^{\lambda_i}} - I_{y_i>0}\right]\frac{w_i}{1+w_i}z_{it}\right\} = 0$$

where $w_i = \frac{\phi_i}{1-\phi_i} = \exp\{\sum_{j=1}^{m} z_{ij}\delta_j\}$. Furthermore, the expression for $\partial \sum_{j=y_i}^{\infty} \Pr(y_i = j)/\partial \beta_r$ can be written as follows,

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

2.5 Goodeness-of-fit

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

$$D = -2\left[\log L\left(\hat{\theta}_i;\hat{\lambda}_i\right) - \log L\left(\hat{\theta}_i;y_i\right)\right] \qquad (6)$$

where $\log L\left(\hat{\theta}_i;\hat{\lambda}_i\right)$ and $\log L\left(\hat{\theta}_i;y_i\right)$ are the model's likelihood evaluated respectively under $\hat{\theta}_i$ and y_i (Lambert, 1989).

For an adequate model, the asymptotic distribution of the deviance statistic D is chi-square distribution with $n-k-1$ degrees of freedom which k is the number of estimated parameters. 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 $\hat{\mu}_i$'s may not be reasonably large and so the deviance statistic D may not be suitable. Thus, the log-likelihood statistic $\log L\left(\hat{\theta}_i;y_i\right)$ 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.

When there are several maximum likelihood models, one can compare the performance of alternative models based on several likelihood measures which have been proposed in the statistical literature. The AIC is the most regularly used measure. The AIC is defined as

$$AIC = -2l + 2p$$

where l denotes the log likelihood evaluated under μ and p is the number of parameters. For this measure, the smaller the AIC, the better the model is (Lambert, 1989).

3.0 RESULTS AND DISCUSSION

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 ($camper$).

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 has two values, zero and one as Table 2.

Table 1 Descriptive statistics

Variable	Mean	Std Dev	Min	Max	Variance
Count	3.296	11.635028	0	149	135.373879
Child	0.684	0.850315	0	3	0.7230361
Persons	2.528	1.112730	1	4	1.2381687

Table 2 Camper variable

Camper	Frequency	Percent
0	103	41.2
1	147	58.8

We have considered the model as follow

$$\log \lambda = b_0 + b_1 camper + b_2 persons + b_3 child\,,$$

$$\text{logit}\,\phi_i \text{ or } \text{logit}\,w_0 = a_0 + a_1 camper + a_2\, persons + a_3\, child$$

Furthermore, we put two censoring points, $c_1 = 3, c_2 = 6$. Table 3 shows the estimation of the parameters according to different censoring constants. Also, the $-2LL$ and AIC are presented as the goodness-of-fit measures.

According to the censoring points, there is 22.8% censored data when $c_1 = 3$. It means that 22.8% of the values of the response variable ($count$) is 0,1,2,3 and the rest (77.2%) of the values of the response variable is greater than 3 that is censored in the model. Also the percentage of the censoring for $c_2 = 6$ is 11.6%.

4.0 CONCLUSION

Table 3 shows the parameter estimation, standard errors and goodness-of-fit measures of the CZIP and CHP models for the different censoring points. The standard errors of all models are very close to each other and almost small. Also, the estimated value of each variable for all models are almost similar and it means that all regression models show quite a similar effect of each variable on the data. For instance, the estimated camper of the ZIP and HP models is a positive value for the different

censoring points, it means that while being a camper ($camper = 1$), the expected $\log(count)$ will be increased by 0.3843; 0.4247; 0.3927 and 0.4247 respectively for ZIP ($c_1 = 3$), HP ($c_1 = 3$), ZIP ($c_2 = 6$) and HP ($c_2 = 6$).

Table 3 Estimated ZIP vs HP with censoring. (Numbers in parenthesis are standard error of the estimates)

	$c_1 = 3$		$c_2 = 6$	
Parameter	**CZIP**	**CHP**	**CZIP**	**CHP**
a_0	1.5321 (0.5876)	2.3087 (0.4612)	1.7449 (0.5267)	2.3087 (0.4612)
a_1	-0.952 (0.4136)	-1.0179 (0.3246)	-0.9356 (0.3681)	-1.0179 (0.3246)
a_2	-0.9613 (0.2269)	-1.1104 (0.1911)	-0.9821 (0.2083)	-1.1104 (0.1911)
a_3	2.1395 (0.3673)	2.138 (0.3107)	2.0655 (0.3366)	2.138 (0.3107)
b_0	-0.4261 (0.2902)	-0.4326 (0.2949)	-0.241 (0.222)	-0.4326 (0.2949)
b_1	0.3843 (0.1884)	0.4247 (0.1895)	0.3927 (0.1397)	0.4247 (0.1895)
b_2	0.4532 (0.0859)	0.4495 (0.0864)	0.471 (0.0624)	0.4495 (0.0864)
b_3	-0.4641 (0.1759)	-0.4762 (0.1745)	-0.557 (0.1296)	-0.4762 (0.1745)
$-2LL$	440.3	440.4	584.4	584
AIC	456.3	456.4	600.4	600

The goodness-of-fit measure of the CZIP model is very close to the CHP model for each censoring point. As expected, the goodness-of-fit measure of the CZIP and CHP models increases when the percentage of censoring decreases (or the value of the censoring point increases) and it is because of the used data in the noncensored part of the regression model.

Table 4 shows the estimated count variable of the zero-inflated and hurdle model versus the real count of the Fish data when the censoring point is 3 and 6. In this case, the number of zeros in the real data is 142 and the censored hurdle Poisson regression model estimate 142 zeros too for both censoring points. Furthermore, the CHP regression model show a closer estimates to the real counts compared to CZIP regression model for the number of ones to the number of censored values for both censoring points.

Table 4 Estimated CZIP and CHP models vs real count

First censoring point				Second censoring point			
count	**Real**	**CZIP**	**CHP**	***count***	**Real**	**CZIP**	**CHP**
0	142	139.87	142	**0**	142	139.87	142
1	31	31.82	31.04	**1**	31	23.4	22.94
2	20	33.75	33.04	**2**	20	30.11	29.53
Cen	57	44.57	43.92	**3**	12	25.83	25.33
				4	6	16.62	16.3
				5	10	8.55	8.39
				Cen	29	5.62	5.51

Acknowledgement

We would like to acknowledge the financial support from Universiti Teknologi Malaysia for the Research University Grant.

References

[1]. Caudill, S. B. and Mixon Jr., F. G. 1995. Modeling Household Fertility Decisions: Estimation and Testing of Censored Regression Models for Count Data. *Empirical Economics*. 20(2): 183–197.
[2]. Greene, W. 2007. Functional Form and Heterogeneity in Models for Count Data. *Foundations and Trends* in Econnometrics. 1(2): 113–218.
[3]. Gurmu, S. 1998. Generalized Hurdle Count Data Regression Models. *Economics Letters*. 58: 263–268.
[4]. Lambert, P. J. 1989. The Distribution and Redistribution of Income - A Mathematical Analysis. Oxford, U.K.: Basil Blackwell.
[5]. Lambert, P. J. 1992. Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing. Technometrics. 34(1): 1–14.
[6]. Mullahy, J. 1986. Specication and Testing of some Modified Count Data Models. *Journal of Econometrics*. 33: 341–365.
[7]. Terza, J. V. 1985. A Tobit-Type Estimator for the Censored Poisson Regression Model. *Economics Letters*. 18: 361–365.