Jurnal Teknologi, 57 (Sciences & Engineering) Suppl 1, March 2012: 189–198
© Penerbit UTM Press, Universiti Teknologi Malaysia

PARAMETER ESTIMATION ON HURDLE POISSON REGRESSION MODEL WITH CENSORED DATA

SEYED EHSAN SAFFARI[1*], ROBIAH ADNAN[2]
& WILLIAM GREENE[3]

Abstract. A Poisson model typically is assumed for count data. In many cases, there are many zeros in the dependent variable and because of these many zeros, the mean and the variance values of the dependent variable are not the same as before. In fact, the variance value of the dependent variable will be much more than the mean value of the dependent variable and this is called over-dispersion. Therefore, Poisson model is not suitable anymore for this kind of data because of too many zeros. Thus, it is suggested to use a hurdle Poisson regression model to overcome over-dispersion problem. Furthermore, the response variable in such cases is censored for some values. In this paper, a censored hurdle Poisson regression model is introduced on count data with many zeros. 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 on estimated parameters and their standard errors via an example.

Keywords: Hurdle Poisson regression; censored data; maximum likelihood method; goodness-of-fit

1.0 INTRODUCTION

Commonly, the starting point for modeling the number of reported claims is the Poisson distribution with the probability function is given as:

$$f_Y(y) = \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!} \tag{1}$$

[1,2] Department of Mathematical Sciences, Faculty of Science, Universiti Teknologi Malaysia, 81310 UTM Johor Bahru, Malaysia

[3] 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

where covariates are included in the model through the parameter $\lambda_i = \exp(x_i'\beta)$ (Dionne and Vanasse, 1989). The Poisson distribution is equidispersed since its mean and variance are both equal to λ_i. The distribution has some severe drawbacks that limit its use, hence other distributions can be used, such as hurdle models (Boucher *et al.*, 2007).

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.

The most important usage of a hurdle count data model is the hurdle at zero. The hurdle at zero formulation can account for excess zeros. It means that this model can be used at the situations where there are many zeros at the response variable. In this case, the hurdle at zero defines a probability ($\Pr(Y = 0)$) that is the first part of the two-part models.

The hurdle model is flexible and can handle both under- and overdispersion problems. A generalized hurdle model is introduced by Gurmu (1998) for the analysis of overdispersed or underdispersed count data. Greene (2005) has discussed the comparison between hurdle and zero-inflated models as two part-models. Some researchers have discussed the applications of hurdle models, such as Pohlmeier and Ulrich (1995), Arulampalam and Booth (1997). A hurdle model to the annual number of recreational boating trips by a family is discussed by Gurmu and Trivedi (1996). Dalrymple *et al.* (2003) applied three mixture models including a hurdle model and argued its application in the incidence of sudden infant death syndrome (SIDS). Boucher *et al.* (2007) compared generalized heterogeneous, zero-inflated, hurdle, and compound frequency models for the annual number of claims reported to the insurer.

Suppose that $g_1(0)$ is the probability value when the value for response variable is zero and that $g_2(k), k = 1, 2, \dots$ is a probability function when the response variable is a positive integer. Therefore, the probability function of the hurdle-at-zero model is given by:

$$P(Y_i = k) = \begin{cases} g_1(0) & \text{for } k = 0 \\ (1 - g_1(0))g_2(k) & \text{for } k = 1, 2, \dots \end{cases} \tag{2}$$

Mullahy (1986) discussed the hurdle-at-zero model and he believes that both parts of the hurdle model are based on probability functions for nonnegative integers such af_1s and f_2. In terms of the general model above, let $g_1(0) = f_1(0)$ and $g_2(k) = f_2(k)/(1 - f_2(0))$. In the case of g_2, normalization is required because f_2 has support over the nonnegative integers $(k = 0,1,\ldots)$ whereas the support of g_2 must be over the positive integers $(k = 1,2,\ldots)$. This means that we need to truncate the probability function f_2. However, this is a theoretical concept, i.e., truncation on f_2 does not mean that there is truncation of the population. All we need to do is to work with a distribution with positive support, and the second part of a hurdle model can use a displaced distribution or any distribution with positive support as well.

Under the Mullahy (1986) assumptions, the probability distribution of the hurdle-at-zero model is given by

$$f(y = 0) = f_1(0)$$

$$f(y = k) = \frac{1 - f_1(0)}{1 - f_2(0)} f_2(k) = \theta f_2(k)\,, k = 1,2,\ldots$$

where f_2 is referred to as parent-process. The numerator of θ presents the probability of crossing the hurdle and the denominator gives a normalization that accounts for the (purely technical) truncation of f_2. It follows that if $f_1 = f_2$ or, equivalently, $\theta = 1$ then the hurdle model collapses to the parent model. The expected value of the hurdle model is given by

$$E(Y) = \theta \sum_{k=1}^{\infty} k f_2(k)$$

and the difference between this expected value and the expected value of the parent model is the factor θ. In addition, the variance value of the hurdle model is given by

$$Var(Y_i) = \theta \sum_{k=1}^{\infty} k^2 f_2(k) - \left[\theta \sum_{k=1}^{\infty} k f_2(k)\right]^2$$

If θ exceeds 1, it means that the probability of crossing the hurdle is greater than the sum of the probabilities of positive outcomes in the parent model. Therefore, increasing the expected value of the hurdle model is related to the expected value of the parent model. Alternatively, if θ is less than 1 (that is the usual case in an application with excess zeros), it means that the probability of not crossing the hurdle is greater than the probability of a zero in the parent model, thus decreasing the expected value of the hurdle model relatively to the expected value of the parent model. Therefore, this model gives a new explanation of excess zeros as being a characteristic of the mean function rather than a characteristic of the variance function.

Consequently, the model can be overdispersed and that depends on the value of the parent processes. To overcome over-dispersion, we would like to censor some values of the response variable that are very big. In statistics, this is called censoring and because we want to censor the values that are bigger than a constant, it is called a right censoring.

The interesting point of the hurdle model is to estimate the parameters by two separate steps. In fact, we can estimate the zero-part parameters by using MLE on the first part of the likelihood function while the other parameters only use the second part which is composed of only non-zero elements. We have used SAS codes for the application part and this characteristic of the model helps us to computationally save time in estimation process.

In this article, the main objective is to explain how we can use hurdle Poisson regression model in right censored data. In section 2, the hurdle Poisson regression model is defined and the likelihood function of hurdle Poisson 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 regression model in right censored data is proposed. An example is conducted for a censored hurdle Poisson regression model in terms of the parameter estimation, standard errors and goodness-of-fit statistic in section 5.

2.0 THE MODEL

Let $Y_i, i = 1,2, \dots, n$ be a nonnegative integer-valued random variable and suppose $Y_i = 0$ is observed with a frequency significantly higher than that which can be modeled by the usual model. We consider a hurdle Poisson regression model in which the response variable $Y_i (i = 1, \dots, n)$ has the distribution

$$P(Y_i = y_i) = \begin{cases} w_0 & ; \quad y_i = 0 \\ (1 - w_0)\dfrac{e^{-\lambda_i} \lambda_i^{y_i}}{(1 - e^{-\lambda})y_i!}; & y_i > 0 \end{cases} \tag{3}$$

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

$$\text{logit}(w_0) = \log\left(\frac{w_0}{1-w_0}\right) = \sum_{j=1}^{m} z_{ij}\delta_j \tag{4}$$

where $z_i = (z_{i1} = 1, z_{i2}, \ldots, z_{im})$ 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 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

$$\log(\lambda_i) = \sum_{j=1}^{m} x_{ij}\beta_j \tag{5}$$

and β_j's are the independent variables in the regression model and m is the number of these independent variables. Furthermore, in this paper we suppose that w_0 and θ_i are not related.

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

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

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

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

We now calculate the log-likelihood function for the hurdle Poisson regression model, we have

$$LL = \sum_{i=1}^{n} \left\{ (1 - d_i)\left[I_{y_i=0} \log w_0 \right. \right.$$
$$+ I_{y_i>0}\{\log(1 - w_0) - \lambda_i + \mathrm{y_i} \log \lambda_\mathrm{i} - \log(\mathrm{y_{i!}})$$
$$\left. \left. - \log(1 - \mathrm{e}^{-\lambda_\mathrm{i}})\} \right] + d_i \log \sum_{j=y_i}^{\infty} Pr(Y_i = j) \right\} \qquad (6)$$

3.0 PARAMETER ESTIMATION

In this section, we obtain the estimated parameters 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

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

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

where

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

4.0 GOODNESS-OF-FIT STATISTICS

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

$$D = -2\left[\log L(\hat{\theta}_i; \hat{\lambda}_i) - \log L(\hat{\theta}_i; y_i) \right] \qquad (7)$$

where $\log L(\hat{\theta}_i; \hat{\lambda}_i)$ and $\log L(\hat{\theta}_i; y_i)$ are the model's likelihood evaluated respectively under $\hat{\theta}_i$ 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 $\hat{\mu}_i$'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 Akaike Information Criterion (AIC) is the most regularly used measure, defined as

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

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

5.0 AN APPLICATION

The state wildlife biologists are interested to model how many fish are being caught by fishermen at a state park[1]. 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 ($camper$).

[1] The fish dataset is available at the UCLA Academic Technology Services website, http://www.ats.ucla.edu.

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 follows

$$\log\lambda = b_0 + b_1 camper + b_2 persons + b_3 child\,,$$
$$\operatorname{logit} w_0 = a_0 + \ a_1 child$$

Furthermore, we put two censoring points, $c_1 = 3, c_2 = 5$. Table 3 shows the estimation of the parameters and the corresponding standard errors (in bracket) 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 $c_2 = 5$ is 12%.

6.0 CONCLUSION

In this article we want to show that the hurdle Poisson regression model can be used to fit right censored data. The hurdle Poisson regression model with right censoring is fitted to these real data. The results from the fish data are summarized in Table 1-3. The goodness-of-fit measures are presented in the Table 3 according to different censoring points and it is obvious that we have a smaller value for $-2\mathrm{LL}$ or AIC when the percentage of censoring increase and that is because of the number of the data which are used in the model.

Table 3 Parameter estimation

Parameter	**22.8 % censored**	**12 % censored**
$\mathbf{b_0}$	−0.3682 (0.2589)	−0.2541 (0.2237)
$\mathbf{b_1}$	0.4377 (0.1660)	0.4083 (0.1396)
$\mathbf{b_2}$	0.4647 (0.0753)	0.4728 (0.0623)
$\mathbf{b_3}$	−0.5493 (0.1550)	−0.5621 (0.1285)
$\mathbf{a_0}$	−0.3843 (0.1703)	−0.3843 (0.1703)
$\mathbf{a_1}$	1.1110 (0.2049)	1.1110 (0.2049)
−2LL	547.1	635.3
AIC	559.1	647.3

ACKNOWLEDGEMENTS

We would like to acknowledge the financial support from Universiti Teknologi Malaysia for the Research University Grant (Q.J1300000.7126.02J67).

REFERENCES

[1] G. Dionne and C. Vanasse, 1989. A Generalization of Automobile Insurance Rating Models: The Negative Binomial Distribution with Regression Component. *Astin Bulletin*. 19: 199-212.

[2] J. P. Boucher, M. Denuit and M. Guill´en, 2007. Risk Classification for Claim Counts: A Comparative Analysis of Various Zero-Inflated Mixed Poisson and Hurdle Models. *North American Actuarial Journal*. In Press.

[3] J. Mullahy, 1986. Specification and Testing of Some Modified Count Data Models. *Journal of Econometrics*. 33: 341-365.

[4] S. Gurmu, 1998. Generalized Hurdle Count Data Regression Models. *Economics Letters*. 58: 263-268.

[5] W. Greene, 2005. Functional Form and Heterogeneity in Models for Count Data. *Foundations and Trends ® in Econometrics*. 1(2): 113-218.

[6] W. Pohlmeier and V. Ulrich, 1995. An Econometric Model of the Two-Part Decision-Making Process in the Demand for Health Care. *The Journal of Human Resources.* 30: 339-361.

[7] W. Arulampalam and A. Booth, 1997. Who Gets Over the Training Hurdle? A study of the Training Experiences of Young Men and Women in Britain. *Journal of Population Econometrics*. 10: 197-217.

[8] S. Gurmu and P. K. Trivedi, 1996. Excess Zeros in Count Models for Recreational Trips. *Journal of Business and Economic Statistics*. 14: 469–477.

[9] M. Dalrymple, I. Hudson and A. Barnett, 2003. Finite Mixture, Zero-inflated Poisson and Hurdle models with application to SIDS. *Computational Statistics & Data Analysis*. 41: 491-504.

[10] S. E. Saffari and Robiah Adnan, 2011. Zero-Inflated Poisson Regression Models with Right Censored Count Data. *Matematika*. 27(1): 21-29.

[11] S. E. Saffari, Robiah Adnan and W. Greene, 2011. Handling of Over-Dispersion of Count Data via Truncation using Poisson Regression Model. *Journal of Computer Science & Computational Mathematics*. 1(1): 1-4.

[12] S. E. Saffari and Robiah Adnan, 2011. Zero-Inflated Negative Binomial Regression Model with Right Censoring Count Data. *Journal of Materials Science and Engineering*. B1: 551-554.