Handling of Over-Dispersion of Count Data via Truncation using Poisson Regression Model

Seyed Ehsan Saffari[1], Robiah Adnan[2] and William Greene[3]
[1,2] Universiti Teknologi Malaysia, Mathematics Department, Faculty of Science,
[3] Department of Economics, Stern School of Business, New York University
ehsanreiki@yahoo.com, robiaha@utm.my, wgreene@stern.nyu.edu

***Abstract*:** A Poisson model typically is assumed for count data. It is assumed to have the same value for expectation and variance in a Poisson distribution, but most of the time there is over-dispersion in the model. Furthermore, the response variable in such cases is truncated for some outliers or large values. In this paper, a Poisson regression model is introduced on truncated 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 truncation in terms of parameters estimation and their standard errors via real data.

***Keywords*:** Poisson regression, over-dispersion, truncation, parameter estimation.

1. Introduction

There are many statistical applications, when the random variable Y represents counts. Examples of count data include the number of students drop out, the number of failures of an experiment per unit time, or the number of accidents on a highway per unit time. There are many studies dealing with count data and various distributions have been proposed for 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.

$$P(Y = y) = \frac{e^{-\lambda}\lambda^{y}}{y!}, \qquad y = 0,1,2,\dots$$

$$E(Y) = Var(Y) = \lambda$$

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 (Cameron and Trivedi, 1998). Another type of over-dispersion relative to Poisson distribution is that in such cases there are some outliers or some large values which have effects on the variance and mean values. In this case, the variance value is greater than mean value and therefore we have over-dispersion in the model. To overcome over-dispersion, we would like to cut the values of the response variable that are very big. In statistics, this is called truncation and because we want to truncate the values that are bigger than a constant, it is called a right truncation.

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

2. The Model

Let $Y_i, i = 1,2,3,\dots,n$ be a nonnegative integer-valued random variable from a Poisson distribution. Thus, the regression model is defined as

$$P(Y_i = y_i|\boldsymbol{x_i}) = \frac{e^{-\lambda_i}{\lambda_i}^{y_i}}{y_i!}, \qquad y_i = 0,1,2,\dots \quad (1)$$

When there is interest in capturing any systematic variation in λ_i, the value of λ_i is most commonly placed within a loglinear model

$$log(\lambda_i) = \sum_{j=1}^{m} x_{ij}\beta_j \quad (2)$$

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

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 ith 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

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

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

$$log\, L(\theta_i; y_i) = \sum_{i=1}^{k} \big[log\, f(y_i;\theta_i) - log\big(1 - Pr(Y_i \geq y_i)\big) \big] \quad (4)$$

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

$$P_T(Y_i = y_i) = \frac{e^{-\lambda_i} {\lambda_i}^{y_i}}{y_i!\left(1 - \sum_{y_i=t_i+1}^{\infty} P(Y_i = y_i|x_i)\right)} \quad (5)$$

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 Poisson regression model with right truncation as follow

$$LL = \sum_{i=1}^{k} \left[-\lambda_i + y_i\, log\, \lambda_i - log\, y_i! - log\left(1 - \sum_{y_i=t_i+1}^{\infty} P(Y_i = y_i|x_i)\right)\right] \quad (6)$$

where k is the number of observation after truncation.

3. 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

$$\frac{\partial LL}{\partial \beta_r} = \sum_{i=1}^{k} \left[-\lambda_i + y_i + \frac{\sum_{y_i=t_i+1}^{\infty} (y_i - \lambda_i) P(Y_i = y_i|x_i)}{1 - \sum_{y_i=t_i+1}^{\infty} P(Y_i = y_i|x_i)} \right] x_{ir} = 0 \quad (6)$$

4. 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

$$D = -2\big[log\, L(\hat{\theta}_i; \hat{\lambda}_i) - log\, L(\hat{\theta}_i; y_i)\big] \quad (7)$$

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 $\log L(\hat{\theta}_i; y_i)$ 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.
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 the number of parameters. For this measure, the smaller the AIC, the better the model is.

5. 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 ($camper$).
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

Figure 1 shows the histogram of the count variable and it is clear that we have zero-inflation problem, also we have few big values that we are interested to truncate them.

We have considered the model as follow

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

Furthermore, we put two truncation points, $t_1 = 3, t_2 = 5$. 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 $t_1 = 3$. 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 truncated in the model. Also the percentage of the truncation for $t_2 = 5$ is 12%. Furthermore, the values of the independent variables ($camper, persons, child$) are truncated for those values of response variable which is truncated. For example, the 25th value of the response variable is $count_{25} = 30$, and the values of the independent variables are as follow

$$camper_{25} = 1, persons_{25} = 3, child_{25} = 0$$

So we have to truncate these values of the independent variables because the value of their response variable should be truncated ($count_{25} > truncation\ point$).

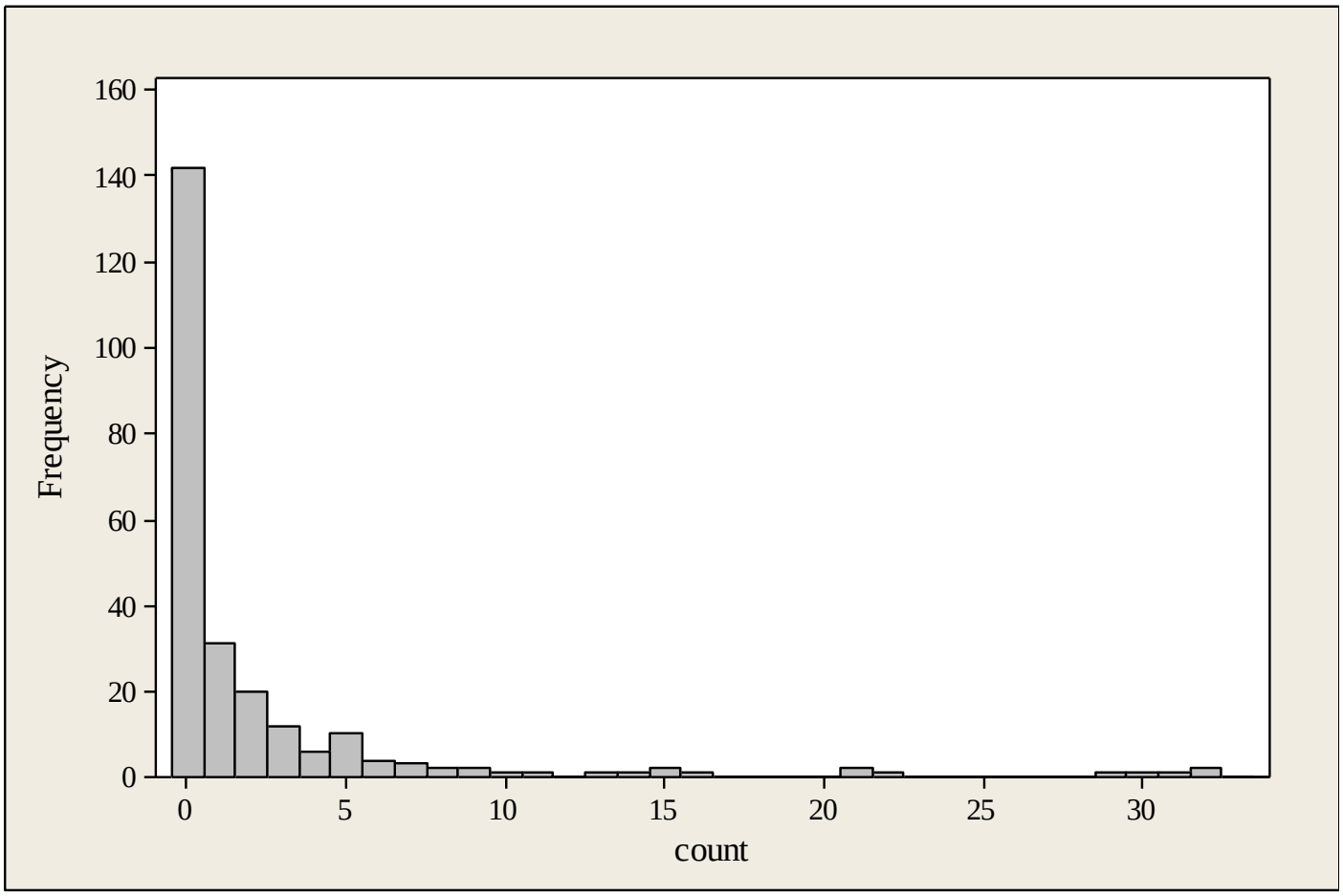

Figure 1: Histogram of the response variable

Table 3: Parameter Estimation

parameter	$t_1 = 3$	$t_2 = 5$
b_0	-1.5642 (0.3153)	-1.7486 (0.2605)
b_1	0.4382 (0.2160)	0.7601 (0.1727)
b_2	0.6343 (0.1269)	0.7245 (0.0883)
b_3	-1.4224 (0.2162)	-1.2621 (0.1462)
$-2LL$	363.7	516.5
AIC	371.7	524.5

6. Conclusion

In this article we want to show that the Poisson regression model can be used to fit right truncated data. The Poisson regression model with right truncation (TPR) 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 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.

Acknowledgment

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

References

[1] A. C. Cameron, P. K. Trivedi, Regression analysis of count data, *Cambridge University Press*, Cambridge, UK, 1998.

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

[3] P. C. Consul, F. Famoye, "Generalized Poisson regression model", *Comm. Statist. Theory Methods*, vol. 21, pp. 81-109, 1992.

[4] D. Lambert, "Zero-inflated Poisson regression, with an application to defects in manufacturing", *Technometric*, vol. 34, pp. 1-14, 1992.

[5] S. E. Saffari, Robiah Adnan, "Zero-inflated negative binomial regression model with right censoring count data", In *Proc Faculty of Science Postgraduate Conference*, Malaysia, 2010.

[6] S. E. Saffari, Robiah Adnan, "Zero-Inflated Poisson Regression Models with Right Censored Count Data", *Matematika*, vol. 27, no. 1, 2011.

[7] F. Famoye, K. P. Singh, "On inflated generalized Poisson regression model", *Advances and Applications in Statistics*, vol. 3, pp. 135-58, 2003.

[8] F. Famoye, J. T. Wulu, K. P. Singh, "On the generalized Poisson regression model with an application to accident data", *Journal of Data Science*, vol. 2, pp. 287-95, 2004.

[9] S. Bae, F. Famoye, J. T. Wulu, A. A. Bartolucci, K. P Singh, "A rich family of generalized Poisson regression models with applications", *Mathematical and Computers in Simulation*, vol. 69(1-2), pp. 4-11, 2005.

[10] F. Famoye, K. P. Singh, "Zero-inflated generalized Poisson regression model", *Journal of Data Science*, vol. 4, pp. 117-30, 2006.