CAUCHY – JURNAL MATEMATIKA MURNI DAN APLIKASI
Volume 4(3) (2016), Pages 115-119

Estimation Parameters and Modelling Zero Inflated Negative Binomial

Cindy Cahyaning Astuti[1], Angga Dwi Mulyanto[2]

[1] Muhammadiyah University of Sidoarjo, Sidoarjo, Indonesia
[2] Alpha Research, Malang, Indonesia

Email: cindy.cahyaning@umsida.ac.id, angga.dwi.m@gmail.com

ABSTRACT

Regression model between predictor variables and the Poisson distributed response variable is called Poisson Regression Model. Since, Poisson Regression requires an equality between mean and variance, it is not appropriate to apply this model on overdispersion. Poisson regression can be used to analyze count data but it has not been able to solve problem of excess zero value on the response variable. An alternative model which is more suitable for overdispersion data and can solve the problem of excess zero value on the response variable is Zero Inflated Negative Binomial (ZINB). In this research, ZINB is applied on the case of Tetanus Neonatorum in East Java. The aim of this research is to examine the likelihood function and to form an algorithm to estimate the parameter of ZINB and also applying ZINB model in the case of Tetanus Neonatorum in East Java. Maximum Likelihood Estimation (MLE) method is used to estimate the parameter on ZINB and the likelihood function is maximized using Expectation Maximization (EM) algorithm. Test results of ZINB regression model showed that the predictor variable have a partial significant effect at negative binomial model is the percentage of pregnant women visits and the percentage of maternal health personnel assisted, while the predictor variables that have a partial significant effect at zero inflation model is the percentage of neonatus visits.

Keywords: Overdispersion, Tetanus Neonatorum, Zero Inflation, Zero Inflated Negative Binomial (ZINB)

INTRODUCTION

Regression analysis is used to determine relationship between one or several response variable (Y) with one or several predictor variables (X). In the classical linear model assumptions are response variables follow a normal distribution, but in fact often found the response variable did not follow the normal distribution. To overcome this there is development in the classical linear model, namely the Generalized Linear Model (GLM) [1]. GLM assuming the response variable follows the exponential family distribution, which has a more general characteristic. In some research, there are often data with response variable that follows a Poisson distribution, regression analysis is used to this kind of data is the Poisson regression analysis. Poisson regression model is commonly used to analyze the data count (data count). Poisson regression there is an assumption on $Y \sim$ Poisson (μ). A key assumption in the Poisson regression analysis is the variance should be equal to the average, the condition is called equidispersion. On the type of count data often encountered zero value is more than 50 percent on the response variable (zero inflation) [2]. Data proportion that has exaggeration zero value can lead to the accuracy of inference. Poisson regression can be used to analyze the data count but still cannot resolve the problem of excessive zero value. In modelling count data if there ar many zero observations on response variable it can be overcome by using Zero inflated Poisson regression (ZIP) model [3]. However, if there are many

Submitted: August, 20 2016 Reviewed: October, 31 2016 Accepted: November, 29 2016

DOI: http://dx.doi.org/10.18860/ca.v4i3.3656

zero observations and occurs overdispersion then Zero inflated Poisson regression (ZIP) inappropriately used. Overdispersion can be defined as a condition in which the Poisson distribution variance is greater than average. If in modelling count data (data count) there are many zero observations on response variable (zero inflation) and occurs overdispersion then the regression model can be used is Zero Inflated Generalized Poisson [2].

In progress there are other alternatives to modelling many zero observations and occurs overdispersion besides using Zero Inflated Generalized Poisson (ZIGP), the regression model is Zero Inflated Negative Binomial (ZINB). Zero Inflated Negative Binomial (ZINB) model is formed of Poisson Gamma mixture distribution [4]. Zero Inflated Negative Binomial (ZINB) can be used as an alternative to modelling many zero observations and occurs overdispersion because this model does not require the variance should be equal with average, in addition Zero inflated Negative Binomial (ZINB) model also has a dispersion parameter that useful to describe the variation of the data, which is commonly denoted by κ (kappa). The purpose of this research is examine the likelihood form, estimation parameters of Zero inflated Negative Binomial (ZINB) model and modelling Zero inflated Negative Binomial (ZINB) on Neonatorum Tetanus cases.

METHODS

In this research used secondary data sourced from East Java Health Profile 2012 [5]. Unit of observation in this research was 38 districts/cities in East Java province which covers 29 districts and 9 Cities. The response variable (Y) used in this research is number of cases of Tetanus Neonatorum in each district/city in East Java province, while the predictor variable (X) is used as much as 4 variables. Operational definition of each variable response and predictor variables will be described as

a. The response variable (Y): Number of cases of Tetanus Neonatorum
b. Predictor variable (X)
   1. The percentage of pregnant mothers visit K4 (X_1)
   2. Percentage of immunization Tetanus Toxoid (TT) in pregnant women (X_2)
   3. Percentage of maternal mothers assisted by health workers (X_3)
   4. The percentage of neonates visits (X_4)

The method of analysis in this study is.

a. Knowing the probability function of Zero inflated Negative Binomial (ZINB) model.
b. Determining the likelihood function of Zero inflated Negative Binomial (ZINB) model based on probability function that are already known.
c. Develop algorithms for estimation parameter process based on the likelihood function that is already known. Parameter estimation of Zero inflated Negative Binomial (ZINB) model. was performed using MLE method and solved using EM algorithm.
d. Modelling Zero inflated Negative Binomial (ZINB) model on Neonatorum Tetanus cases in East Java Province
e. Significance test of parameters model carried out simultan and partial test. Statistical tests are used for simultan test is the test statistic G and to partial test used test statistics t.

RESULTS AND DISCUSSION

Estimation parameter Zero inflated Negative Binomial (ZINB) was conducted using Maximum Likelihood Estimation (MLE) and to maximize the function is used EM (Expectation Maximization) algorithm. Probability Function of Zero inflated Negative Binomial (ZINB) model can be defined as :

$$P(Y_i = y_i) = \begin{cases} \pi_i + (1-\pi_i)\left(\frac{1}{1+\boldsymbol{\kappa}\mu_i}\right)^{\frac{1}{\boldsymbol{\kappa}}}, for\ y_i = 0 \\ (1-\pi_i)\frac{\Gamma\left(y_i+\frac{1}{\boldsymbol{\kappa}}\right)}{\Gamma\left(\frac{1}{\boldsymbol{\kappa}}\right)y_i!}\left(\frac{1}{1+\boldsymbol{\kappa}\mu_i}\right)^{\frac{1}{\boldsymbol{\kappa}}}\left(\frac{\boldsymbol{\kappa}\mu_i}{1+\boldsymbol{\kappa}\mu_i}\right)^{y_i}, for\ y_i > 0 \end{cases}$$

EM algorithm consists of two stage, expectation and maximization stage. Expectation stage is expectation calculation of ln likelihood the function, the next stage maximization is calculation to look for estimation parameter which maximizes the likelihood function. Probability function of ZINB model consist of two conditions, $y_i = 0$ and $y_i > 0$. Response variable is also composed of two conditions, namely zero state and negative binomial state. To describe in detail the condition y_i, then it will be redefined variables y_i with latent variable Zi.

$$Z_i = \begin{cases} 1, if \;\; y_i \; from \; zero \; state \\ 0, if \;\; y_i > 0 from \; negative \; binomial \; state \end{cases}$$

Zero inflated Negative Binomial regression (ZINB) model can be defined as two models that are :
Model for *negative binomial* $\hat{\mu}_i$

$$ln\hat{\mu}_i = \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_0 x_{ij} \, , i = 1,2, \dots, n \; and j = 1,2,..,p$$

Model for *zero inflation* $\hat{\pi}_i$

$$logit\hat{\pi}_i = \hat{\gamma}_0 + \sum_{j=1}^{p} \hat{\gamma}_0 x_{ij} \, , i = 1,2, \dots, n \; and j = 1,2,..,p$$

EM algorithm is alternative methods to maximize likelihood function on the data containing latent variables defining new variables such as variable Zi. EM algorithm consists of two stage: the expectation stage and maximization stage. Expectation stage is calculation of the ln likelihood function, the next stage is maximization calculation stage to look for parameter estimation which maximizes the likelihood function ln results from stage earlier expectations. Estimation parameter and parameter test of Zero inflated Negative Binomial (ZINB) on Neonatorum Tetanus cases in East Java Province using SAS software, the result can see at table 1.

Table 1. Estimation Parameter and Parameter Test of ZINB

Parameter	Estimation	SE	*t* value	(Pr >| *t*|)
$\hat{\beta}_0$	-5,847	3,602	-1,623	0,105
$\hat{\beta}_1$	-0,145	0,055	-2,644	0,008*
$\hat{\beta}_2$	-0,006	0,010	-0,599	0,549
$\hat{\beta}_3$	0,233	0,101	2,295	0,022*
$\hat{\beta}_4$	-0,023	0,067	-0,339	0,735
$\hat{\gamma}_0$	11,325	13,409	0,845	0,398
$\hat{\gamma}_1$	0,223	0,169	1,316	0,188
$\hat{\gamma}_2$	-0,296	0,179	-1,653	0,098
$\hat{\gamma}_3$	0,835	0,503	1,660	0,096
$\hat{\gamma}_4$	-1,078	0,539	-2,000	0,045*
Test Statistic G = 581,24				

Results of simultan parameter test based on test statistic G. G test is 581.24. Value of G test is greater than $\chi^2_{(0,05;8)} = 15,507$. This shows that simultaneously the predictor variables X_1, X_2, X_3 and X_4 have significant effect on response variable. While results of partial parameter test based on test statistical t. According to Table 1, there are two predictor variables in negative binomial state model and one predictor variables in zero inflation state model that has t value greater than or equal to $t_{(\alpha / 2;\, 37 = 2.00)}$ and has p-value less than α (0.05). This indicates that the predictor variables were partial significant effect in negative binomial state model are pregnant mothers visit K4 (X_1) and maternal mothers assisted by health workers (X_3), while the predictor

variables were partial significant effect in zero inflation state model is the percentage of neonates visits (X_4). So that Zero inflated Negative Binomial (ZINB) model can be defined as :

a. Negative binomial state model for $\hat{\mu}$

$$\hat{\mu} = exp(-5{,}847 - 0{,}145\,X_1 - 0{,}006\,X_2 + 0{,}233\,X_3 - 0{,}023\,X_4)$$

b. Zero inflation state model for $\hat{\pi}$

$$\hat{\pi} = \frac{\exp(11{,}325 + 0{,}223\,X_1 - 0{,}296\,X_2 + 0{,}835\,X_3 - 1{,}078\,X_4)}{1 + \exp(11{,}325 + 0{,}223\,X_1 - 0{,}296\,X_2 + 0{,}835\,X_3 - 1{,}078\,X_4)}$$

All coefficient parameter which aren't significant still is exist in Negative binomial state Zero inflation state model because it is intended to determine the contribution of each predictor variable on the response variable can be defined as :

Zero inflation model for $\hat{\pi}$

1. Each additional 1 percent of pregnant mothers visit K4 (X1) it will increase the chances of the number of Tetanus Neonatorum by exp (0.223) = 1.249 times the number of cases of Tetanus Neonatorum original, if the other variables constant value.
2. Each additional 1 percent immunization Tetanus Toxoid (TT) in pregnant women (X2) will decrease the chances of the number of Tetanus Neonatorum by exp (0.296) = 1.344 times the number of cases of Tetanus Neonatorum original, if the other variables constant value.
3. Each additional 1 percent of maternal mothers assisted by health workers (X3) then it will increase the chances of the number of Tetanus Neonatorum by exp (0.835) = 2.305 times the number of cases of Tetanus Neonatorum original, if the other variables constant value.
4. Each additional 1 percent of neonates visits (X4) will decrease the chances of the number of Tetanus Neonatorum by exp (1.078) = 2.939 times the number of cases of Tetanus Neonatorum original, if the other variables constant value.

Let's discussion, based on the negative binomial state model and zero inflation state model, there are signs of regression coefficient as opposed to the theory are percentage of maternal mothers assisted by health workers (X3) to model negative binomial state model and the percentage of pregnant mothers visit K4 (X1) and the percentage of maternal mothers assisted by health workers (X3) for zero inflation state model . The existence of the regression coefficient has a sign contrary to the theory of probability caused by the effect of the multikolinieritas. Moreover sign contrary to the theory also caused by the shape of the data pattern of the predictor variables that have a positive correlation with the response variable. In a subsequent study if there are multikolinieritas the predictor variables can be addressed using Principal Component Analysis (PCA).

CONCLUSION

Based on the results, estimation parameter of Zero inflated Negative Binomial (ZINB) model was conducted using Maximum Likelihood Estimation (MLE) and to maximize the likelihood function used the EM (Expectation Maximization) algorithm. For parameter test predictor variable that has significant effect on the number of cases of Tetanus Neonatorum are are pregnant mothers visit K4 (X_1) and maternal mothers assisted by health workers (X_3) for the negative binomial state models, while zero inflation state model predictor variable that has significant effect on the number of cases of Tetanus Neonatorum include the percentage of neonates visit (X_4).

REFERENCES

[1] A. Agresti, Categorical Data Analysis, New York: John Wiley and Sons, Inc., 2002.

[2] F. Famoye dan K. P. Singh, "Zero Inflated Poisson Regression Model with an Applications Domestic Violence to Accident Data," *Journal of Data Science,* pp. 117-130, 2006.

[3] D. Lambert, "Zero Inflated Poisson Regression, With an Application to Defect in Manufacturing," *Technometric,* vol. 34, no. 1, 1992.

[4] J. M. Hilbe, Negative Binomial Regression, New York: Cambridge University Press, 2011.

[5] Dinas Kesehatan Provinsi Jawa Timur, Profil Kesehatan Provinsi Jawa Timur Tahun 2012, Surabaya: Dinas Kesehatan Provinsi Jawa Timur, 2013.