Modeling Tetanus Neonatorum case using the regression of negative binomial and zero-inflated negative binomial

Luthfatul Amaliana, Umu Sa’adah and Ni Wayan Surya Wardhani
Brawijaya University, Jl. Veteran, Malang, 65145, Indonesia

E-mail: luthfatul@ub.ac.id

Abstract. Tetanus Neonatorum is an infectious disease that can be prevented by immunization. The number of Tetanus Neonatorum cases in East Java Province is the highest in Indonesia until 2015. Tetanus Neonatorum data contain over dispersion and big enough proportion of zero-inflation. Negative Binomial (NB) regression is an alternative method when over dispersion happens in Poisson regression. However, the data containing over dispersion and zero-inflation are more appropriately analyzed by using Zero-Inflated Negative Binomial (ZINB) regression. The purpose of this study are: (1) to model Tetanus Neonatorum cases in East Java Province with 71.05 percent proportion of zero-inflation by using NB and ZINB regression, (2) to obtain the best model. The result of this study indicates that ZINB is better than NB regression with smaller AIC.

1. Introduction
Tetanus Neonatorum (TN) is one of the main causes of infant death in Indonesia. The high infant mortality is mostly caused by unclean birth and cord handling. Tetanus is characterized by a painful muscle stiff. It was an infection that caused by Neurotoxin bacteria produced by Clostridium Tetani on a closed wound. The systematic effort of the government to eliminate TN begins with the TT immunization of pregnant and prospective bride. It followed by School Immunization Month program with DT vaccine and TT vaccine for elementary school children as a long-term control of TN.

The World Health Organization (WHO) together with UNICEF invites developing countries in the world to implement the Elimination of Tetanus Maternal and Neonatal (ETMN) program. Previously, Indonesia was one of the countries with the highest cases in Asia. Based on data from the Ministry of Health in the Health Profile of Indonesia 2015, East Java is the province with the highest TN case in Indonesia, which is 22 cases. Provinces with the second and the third rank of TN cases were Banten and West Kalimantan, with twelve and five cases, respectively. Overall, East Java was also the first province in Indonesia with 663 TN cases [1]. TN data in East Java province showed the proportion of zero-inflation was large (71.05 percent). Therefore, modeling the number of TN cases requires a method that can accommodate zero-inflation problems in the data.

Count data commonly modeled by Poisson regression. However, when over dispersion problem happens in Poisson regression, Negative Binomial regression often becomes an alternative solution. Over dispersion is a condition when the sample mean is smaller than the sample variance. NB distribution is a mixture of Poisson-Gamma distribution [2]. It has a dispersion parameter that capable to overcome the over dispersion problem [3]. However, data on TN cases containing over dispersion and a large proportion of zero-inflation. Thus it needs a method to analyze a large zero-inflation and
over dispersion data. Astuti and Mulyanto [4] using Zero Inflated Negative Binomial regression to model TN case in East Java province in 2012. Another study of ZINB regression is the comparison of two zero-inflation models, which is ZIP and ZINB regression, to obtain the best model with over dispersion problem and 73.67 percent proportion of zero-inflation on TN data in 2013 [5]. The purpose of this study is to model the number of TN cases in East Java province with 71.05 percent proportion of zero-inflation by using NB and ZINB regression. Then, find the best model between NB and ZINB regression.

This paper is structured as follows. Section 2 discusses the literature review about over dispersion, NB and ZINB regression that will be used to model later. Then Section 3 discusses the result of modeling TN case by using NB and ZINB regression.

2. NB and ZINB regression

In this section, both of models that will be used in this study are explained. This study uses two models that have a dispersion parameter to overcome over dispersion problem. However, over dispersion problem will be discussed first.

2.1. Over dispersion

A condition while the sample variance is greater than the sample mean called as over dispersion. This condition is a violation of the equality of mean and variance assumption in Poisson regression. The occurrence of over dispersion can be identified by finding the deviance model and dividing by degrees of freedom model. If the result is greater than 1, then there is over dispersion. Otherwise, if the result is less than 1, then it is said to be under dispersion. The deviance could be written as [6]:

\[ D = 2 \sum_{i=1}^{n} y_i \ln \left( \frac{y_i}{\hat{\mu}_i} \right) \]  

\( n \) is the number of observation, \( y_i \) is dependent variable for \( i = 1,2, \ldots, n \), and \( \hat{\mu}_i \) is mean of dependent variable which is influenced by independent variables.

2.2. NB regression

A generalization of Poisson regression with no restrictive assumption known as Negative Binomial regression. The restrictive assumption of Poisson model is the equality of mean and variance. The traditional NB regression model commonly known as NB-2 is written as [7]:

\[ \ln \hat{\mu}_i = \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j x_{ij} \]  

\( i = 1,2, \ldots, n \), \( j = 1,2, \ldots, p \), \( p \) is the number of independent variables, \( \beta \) is unknown parameter, and a set of \( x \) (independent variables). The parameters of NB regression are estimated using Maximum Likelihood Estimation (MLE) method with Fisher Scoring method to maximize the ln-likelihood function.

The likelihood function of NB model could be written as:

\[ P(Y_i = y_i) = \frac{\Gamma\left(y_i + \frac{1}{\kappa}\right)}{\Gamma\left(\frac{1}{\kappa}\right)\Gamma\left(y_i\right)} \left(1 + \frac{\kappa}{\mu_i}\right)^{-\frac{1}{\kappa}} \left(\frac{\kappa \mu_i}{1 + \kappa \mu_i}\right)^{y_i} \]  

\( \Gamma(.) \) is gamma function, \( \kappa \) is dispersion parameter, and \( Y_i \sim NB(\mu_i, \kappa) \).

2.3. ZINB regression

Zero-inflated negative binomial regression is a model formed from a Poisson-Gamma mixture distribution. The ZINB model can modeling a count data contains zero-inflation in the dependent
variable and over dispersion [8]. The observation \(y_1, y_2, \ldots, y_n\) from dependent variable occured in this two condition:

\[
y_i = \begin{cases} 
0, & \text{with probability } \pi_i \\
\text{NB}(\mu_i, \kappa), & \text{with probability } (1 - \pi_i)
\end{cases}
\]  

(4)

The first condition is called as zero-inflation state. It happened when the dependent variable is zero, while the probability is \(\pi_i\). The other condition called as NB state and the probability is \((1 - \pi_i)\). The likelihood function of ZINB model could be expressed as:

\[
P(Y_i = y_i) = \begin{cases} 
\pi_i + (1 - \pi_i) \left( \frac{1}{1 + \kappa \mu_i} \right)^{\frac{1}{\kappa}}, & \text{for } y_i = 0 \\
(1 - \pi_i) \frac{\Gamma(y_i + \frac{1}{\kappa})}{\Gamma(\frac{1}{\kappa})} \left( \frac{1}{1 + \kappa \mu_i} \right)^{\frac{1}{\kappa}} \frac{1}{1 + \kappa \mu_i}^{\gamma y_i}, & \text{for } y_i > 0
\end{cases}
\]  

(5)

\(0 \leq \pi_i \leq 1, \mu_i \geq 0, \kappa \) is dispersion parameter, while \(\Gamma(\cdot)\) is gamma function. Random variable \(Y_i \sim \text{NB}(\mu_i, \kappa)\) for \(\pi_i = 0\). \(\mu_i\) and \(\pi_i\) are assumed depend on independent variable \(x_i\) that is defined by:

\[
\mu_i = \exp(x_i^T \beta)
\]  

(6)

\[
\pi_i = \frac{\exp(x_i^T y)}{1 + \exp(x_i^T y)} \quad \text{and} \quad (1 - \pi_i) = \frac{1}{1 + \exp(x_i^T y)}
\]  

(7)

The ZINB model for Negative Binomial state \(\hat{\mu}_i\) and zero-inflation state \(\hat{\pi}_i\) for \(i = 1, 2, \ldots, n\) and \(j = 1, 2, \ldots, p\):

\[
\ln \hat{\mu}_i = \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j x_{ij}
\]  

(8)

\[
\logit \hat{\pi}_i = \hat{\gamma}_0 + \sum_{j=1}^{p} \hat{\gamma}_j x_{ij}
\]  

(9)

\(\beta\) and \(\gamma\) are parameters of ZINB model. Parameters of ZINB model is estimated by MLE completed with EM (Expectation-Maximization) algorithm. The maximum ln-likelihood function is:

\[
\ln L(\kappa, \beta, \gamma) = \begin{cases} 
\sum_{i=1}^{n} \ln \left( e^{(x_i^T y)} \frac{1}{1 + e^{(x_i^T y)}} \right)^{\frac{1}{\kappa}}, & \text{for } y_i = 0 \\
\sum_{i=1}^{n} \ln \left( \frac{\Gamma(y_i + \frac{1}{\kappa})}{\Gamma(\frac{1}{\kappa})} \frac{1}{1 + \kappa e^{(x_i^T y)}}^{\gamma y_i}, & \text{for } y_i > 0
\end{cases}
\]  

(10)

3. Method

This study using secondary data based on East Java Province Health Profile 2015 which is published by East Java Province Health Department in 2016 [9]. The data is a cross-sectional data including 38 regency/ city in East Java consist of 29 regencies and 9 cities. The dependant variable is the number of Tetanus Neonatorum (TN) case in every regency/city. The steps of analysis in this study can be described as follows:

1. Do the descriptive analysis of variables
2. Identify proportion of zero-inflation
3. Identify over dispersion
4. Identify multicollinearity
5. Modeling TN case using NB regression
6. Modeling TN case using ZINB regression
7. Testing the significance parameter of NB and ZINB model simultaneously and partially
8. Decide the best model

4. Result and discussion
Data on TN case in every regency/city is a non-negative integer ranging from zero to eight in the sample. Numbers of TN case in East Java is 22 cases from 11 regency/city.

![Histogram of The Number of Tetanus Neonatorum in East Java Province](image)

Figure 1. Histogram of Numbers of TN cases

Figure 1 shows frequency distribution of numbers of TN cases. The graph indicates a highly frequency of zeros and long right tailed. The graph also gives the evidence of over dispersion in data.

The independent variables in this study are percent of TT2+ immunization for pregnant women \((X_1)\), maternity mothers assisted by health personnels in health care facility \((X_2)\), K1 visit of pregnant women \((X_3)\), TT1 immunization of pregnant women \((X_4)\), and TT immunization of fertile age woman \((X_5)\).

| Variable | Mean | Standard Deviation | Min | Max |
|----------|------|--------------------|-----|-----|
| \(Y\)    | 0.579| 1.464              | 0   | 8   |
| \(X_1\)  | 47.720| 52.980             | 0   | 281.90 |
| \(X_2\)  | 95.048| 3.849              | 86.150 | 100.65 |
| \(X_3\)  | 99.724| 4.004              | 91.900 | 109.40 |
| \(X_4\)  | 2.910| 6.280              | 0   | 32.33 |
| \(X_5\)  | 26.310| 46.740             | 0   | 257.71 |

Table 1. Descriptive analysis of variables

Table 1 summarize the descriptive statistics of variables consist of the dependent and independent variables. The highest TN cases occur in Bangkalan regency with eight cases, while TN cases not occur in 27 regencies/cities.
Table 2. Zero-inflation checking

| The number of TN case | Frequency | Percent | Cumulative Percentage |
|-----------------------|-----------|---------|-----------------------|
| 0                     | 27        | 71.05%  | 71.05%                |
| 1                     | 8         | 21.05%  | 92.11%                |
| 2                     | 1         | 2.63%   | 94.74%                |
| 4                     | 1         | 2.63%   | 97.37%                |
| 8                     | 1         | 2.63%   | 100.00%               |

The result of zero-inflation checking on numbers of TN case in East Java is summarized in Table 2. The sample data has a great proportion of zero-inflation, about 71.05 percent.

Table 3. Over dispersion checking

| Deviance | Degrees of Freedom | Deviance / Degrees of Freedom |
|----------|--------------------|-------------------------------|
| 38.13502 | 32                 | 1.192                         |

Table 3 summarizes the result of over dispersion checking. The result of deviance divided by degrees of freedom (df) which is greater than 1 indicates the evidence of over dispersion in the data. It is also indicated by the sample variance of dependent variable that greater than the sample mean in Table 1.

It will be identified in advance whether there is multicollinearity problem among independent variables, before conduct to further analysis. There are several ways to identify multicollinearity and one of them is by looking at the value of Variance Inflation Factor (VIF).

Table 4. Multicollinearity identification

| Variable | X_1 | X_2 | X_3 | X_4 | X_5 |
|----------|-----|-----|-----|-----|-----|
| VIF      | 1.597 | 6.723 | 7.367 | 1.789 | 1.341 |

Table 4 shows that all the predictor variables have a VIF value < 10 [10], so it could be said that there are no multicollinearity among five independent variables. Therefore, all the independent variables can be used to modeling numbers of TN cases by using NB and ZINB regression.
Table 5. Parameter estimation of NB regression

| Parameter | NB regression |  |
|-----------|---------------|---|
|           | Estimate | S.E. | Z-value | p-value |
| $\hat{\beta}_0$ | -13.628 | 9.477 | -1.438 | 0.15043 |
| $\hat{\beta}_1$ | -0.008 | 0.009 | -0.896 | 0.37038 |
| $\hat{\beta}_2$ | 0.210 | 0.171 | 1.234 | 0.21736 |
| $\hat{\beta}_3$ | -0.075 | 0.176 | -0.425 | 0.67058 |
| $\hat{\beta}_4$ | 0.128 | 0.046 | 2.784 | < 0.010* |
| $\hat{\beta}_5$ | 0.0003 | 0.007 | 0.047 | 0.96222 |

Kappa = 1.248
Log-likelihood = -26.190; df = 32
AIC = 75.863

*significant at $\alpha = 0.05$

Table 6. Parameter estimation of ZINB regression

| Parameter | ZINB regression |  |
|-----------|-----------------|---|
|           | Estimate | S.E. | Z-value | p-value |
| $\hat{\beta}_0$ | 3.051 | 9.881 | 0.309 | 0.7575 |
| $\hat{\beta}_1$ | -0.028 | 0.016 | -1.738 | 0.0821** |
| $\hat{\beta}_2$ | 0.265 | 0.095 | 2.800 | < 0.010* |
| $\hat{\beta}_3$ | -0.275 | 0.100 | -2.753 | < 0.010* |
| $\hat{\beta}_4$ | 0.117 | 0.033 | 3.564 | < 0.001* |
| $\hat{\beta}_5$ | 0.009 | 0.008 | 1.124 | 0.2608 |
| $\hat{\gamma}_0$ | 31.620 | 26.527 | 1.192 | 0.2333 |
| $\hat{\gamma}_1$ | -0.072 | 0.047 | -1.530 | 0.1261 |
| $\hat{\gamma}_2$ | 1.139 | 0.824 | 1.381 | 0.1672 |
| $\hat{\gamma}_3$ | -1.374 | 0.822 | -1.672 | 0.0945** |
| $\hat{\gamma}_4$ | -0.645 | 0.413 | -1.560 | 0.1187 |
| $\hat{\gamma}_5$ | 0.045 | 0.028 | 1.622 | 0.1048 |

Log-Kappa = 15.5308
Log-likelihood = -24.00; df = 13
AIC = 74.00883

*significant at $\alpha = 0.05$

**significant at $\alpha = 0.10$
Table 5 and Table 6 summarize the results of NB and ZINB regression using software R 3.3.0. According to Table 5, the number of TN cases was found to be associated only with TT1 immunization of pregnant women ($X_4$). While it was associated with maternity mothers assisted by health personnel in health care facility ($X_2$), K1 visit of pregnant women ($X_3$), and TT1 immunization of pregnant women ($X_4$) according to ZINB regression on Table 6 (at significant level $\alpha = 0.05$) for NB state model. Even TT2+ immunization of pregnant women ($X_1$) for NB state model and K1 visit of pregnant women ($X_3$) for zero-inflation state model also significant at level $\alpha = 0.10$.

According to Table 5, the significance test of NB parameters estimate simultaneously ($\alpha = 0.05$) using G test statistic, $G = -2 \text{Log-likelihood} = 61.863$ and $\chi^2$-table with df = 32 is 46.19426. Thus, it could be concluded that simultaneously all the independent variables in NB regression have a significant influence on numbers of TN cases in East Java Province in 2015. While based on Table 6 the significance test of ZINB parameters estimate simultaneously ($\alpha = 0.05$) using G test statistics, $G = -2 \text{Log-likelihood} = 48.00$ and $\chi^2$-table with df = 13 is 22.36. Therefore, simultaneously all the independent variables in ZINB regression also have a significant influence on the number of TN case in East Java Province in 2015.

This study use AIC (Akaike Information Criterion) value to decide the best between NB and ZINB model. The smaller AIC value of a model shows the better model. The AIC value of NB and ZINB model in Table 5 and Table 6, show that the AIC value of ZINB model is smaller than the AIC value of NB model. This suggests that in this study, the ZINB model is better used to modeling numbers of TN case in East Java Province in 2015.

The Zero-Inflated Negative Binomial model can be expressed as:

- NB state model

$$\hat{\mu} = \exp(3.051 - 0.028X_4 + 0.265X_2 - 0.275X_3 + 0.117X_4 + 0.009X_5) \quad (11)$$

- Zero-inflation state model

$$\hat{\pi} = \frac{\exp(31.620-0.072X_1+1.139X_2-1.374X_3-0.645X_4+0.045X_5)}{1+ \exp(31.620-0.072X_1+1.139X_2-1.374X_3-0.645X_4+0.045X_5)} \quad (12)$$

In zero-inflation state model (12), there is no independent variables that significant at level of significance $\alpha = 0.05$. Therefore, only NB state model is to be interpreted where maternity mothers assisted by health personnel in health care facility ($X_2$), K1 visit of pregnant women ($X_3$), and TT1 immunization of pregnant women ($X_4$) are significant in the model.

The average of TN cases expected will be increased by 1.303 times from the previous cases for a percent increase maternity mothers assisted by health personnel in health care facility, will be decreased by 1.317 times from the previous cases for a percent increase K1 visit of pregnant women, and will be increased by 1.124 times from the previous cases for a percent increase TT1 immunization of pregnant women when other variables are constant.

5. Conclusion

According to the result of modeling numbers of TN cases in East Java Province in 2015, ZINB regression is better than NB regression. The result also indicate that maternity mothers assisted by health personnel in health care facility, K1 visit of pregnant women, and TT1 immunization of pregnant women are associated with numbers of TN cases.

Acknowledgment

This work was supported by BOPTN No. 27/UN10.F09.01/PN/2017 by Faculty of Mathematics and Natural Sciences, Brawijaya University, Malang, Indonesia.
References

[1] Kementerian Kesehatan RI 2016 *Profil Kesehatan Indonesia Tahun 2015* (Jakarta: Kementerian Kesehatan RI)

[2] Hilbe, J M 2011 *Negative Binomial Regression 2nd Edition* (New York: Cambridge University Press)

[3] Amaliana, L 2017 Perbandingan ZIP dan ZINB pada Jumlah Kasus Tetanus Neonatorum di Jawa Timur Tahun 2013 *Prosidng Seminar Nasional Matematika Indo-MS Wilayah Sulawesi* vol 1 (Makassar: Universitas Hasanuddin) 428-38

[4] Astuti, C C, Mulyanto, A D 2016 Pemodelan Regresi ZINB pada Kasus Tetanus Neonatorum di Provinsi Jawa Timur *Cauchy–Jurnal Matematika Murni dan Aplikasi* 4 115

[5] Keswari, N M R, Sumarjaya, I W, and Suciptuwati, N L P 2014 Perbandingan Regresi Binomial Negatif dan Regresi Generalisasi Poisson dalam Mengatasi Overdispersi *E-Jurnal Matematika* 3 107

[6] Agresti, A 2002 *Categorical Data Analysis 2nd Edition* (New York: John Wiley & Sons)

[7] Cameron, A C and Trivedi, P K 2013 *Regression Analysis of Count Data 2nd Edition* (Cambridge: Cambridge University Press)

[8] Garay, A M, Hashimoto, E M, Ortega, E M M, and Lachos, V H 2011 On Estimation and Influence Diagnostics for ZINB Regression Models *Computational Statistics and Data Analysis* 55 1304

[9] Dinas Kesehatan Provinsi Jawa Timur 2016 *Profil Kesehatan Provinsi Jawa Timur Tahun 2015* (Surabaya: Dinas Kesehatan Jawa Timur)

[10] Gujarati, D N 2004 *Basic Econometrics, 4th Edition* (New York: Mc.Graw-Hill Companies)