IOP Conference Series:
Materials Science and
Engineering

PURPOSE-LED
PUBLISHING™

PAPER • OPEN ACCESS

# Simulation on the Zero Inflated Negative Binomial (ZINB) to Model Overdispersed, Poisson Distributed Data

To cite this article: Rahma Fitriani *et al* 2019 *IOP Conf. Ser.: Mater. Sci. Eng.* **546** 052025

View the article online for updates and enhancements.

# Simulation on the Zero Inflated Negative Binomial (ZINB) to Model Overdispersed, Poisson Distributed Data

**Rahma Fitriani**[1]**, Lidia Novita Chrisdiana**[1]**, Achmad Efendi**[1*]

[1)] Department of Statistics, Faculty of Sciences, Universitas Brawijaya, Malang

Corresponding author: a_efendi@ub.ac.id

**Abstract**. Poisson regression analysis shows the relationship between predictor variables and response variables that follow the Poisson distribution which has equal dispersion and average values (λ), a situation called equidispersion. However, the variance can also be greater than the average value, called overdispersion. This can be caused by excess opportunities for the emergence of zero values in the response variable or zero excess. The parameter of the overdispersed data analysis can be underestimated so that the results become biased. This bias issue can be, hopefully, overcome by the Zero Inflated Negative Binomial (ZINB) regression analysis. In the 2016 Maternal Mortality Rate data in Bojonegoro District, overdipersion was overcome by ZINB regression even though there was no significant predictor variable found affecting the response variable. ZINB regression analysis can also be applied to generated data (simulation). We had the data with average $\lambda$ = (0.2, 0.4, 0.6, 0.8, 1.0, 5.0) proportion of zeros $p$ = (0.4, 0.6, 0.8), and the number of observations $n$ = (200,500, 800), with each setting was repeated 100 times. From the simulation study it was found that all overdispersion events were always accompanied by zero excess events but not vice versa. The greater the value of $\lambda$ then the greater the dispersion coefficient. The ZINB regression is proven to be able to overcome overdispersion in various conditions of different values of $\lambda$, $p$, $n$ which can be seen from the value $\tau$ (dispersion coefficient) after ZINB regression is less than 1 in all conditions.

Keywords: Zero Inflated Negative Binomial (ZINB), Overdispersion, Zero Excess, Simulation

## 1. Introduction

Regression analysis is a statistical method used to form a model of the relationship between response variables and predictor variables [1]. If the response variable is continuous, linear regression analysis can be used. But if the response variable is a count then the one that can be used is Poisson regression analysis. Poisson regression analysis indicates the relationship between predictor variables and response variable that follows the Poisson distribution. Poisson distribution is based on the number of events that occur during a time interval in a particular area. One uniqueness of Poisson distribution is that the value of variance and the average of response variable is the same, a situation called equidispersion. But in practice, this is often not fulfilled, for instance with smaller variance value than the expected value of the response variable, underdispersion. Other condition that may occur is that the variance is larger than the expected value of the response variable, overdispersion. The condition is caused by the proportion of zero values appearing in the response variable ($p$) that is excessive or zero excess [2]. This condition can produce a standard error that is smaller than the real one or underestimate [3].

There are several ways to overcome overdispersion due to the probability of zero values appearing in the response variables ($p$), such as Zero Inflated Poisson (ZIP) regression, Zero Inflated Generalized Poisson (ZIGP) regression, and Zero Inflated Negative Binomial (ZINB) regression. Overcoming overdispersion means that the regression analysis method can accommodate overdispersion due to excess zero by modelling overdispersed data without having to eliminate the overdispersion conditions from the data. Based on a conducted research [3], ZINB regression is better in modelling the maternal mortality rate (MMR) in Bali Province in 2014 which experienced overdispersion, compared to ZIP regression. Similar things were also obtained from the research conducted on the mortality rate data for Bojonegoro Regency women in 2016 [4], where overdispersion was also occurred. Based on the aforementioned two studies, the researchers wanted to examine further the ability of the ZINB regression to overcome overdispersion on various characteristics.

In order to find out the ZINB regression capability in modelling data that is overdispersed, generated data were used with various characteristics such as average value (λ), the probability of zero occurrence in the response variable ($p$), and the number of observations ($n$). As a reference in generating data then we used the applied data on maternal mortality in Bojonegoro Regency in 2016 [4]. The maternal mortality rate is defined as a death during pregnancy or within a period of 42 days after the end of pregnancy, due to all causes related to or aggravated by pregnancy or treatment, but not due to accident or injury [7]. The maternal mortality rate is one indicator of the success of health development in Indonesia. This is because mothers and children are vulnerable groups, related to the phase of pregnancy, childbirth, and the stage of growth and development in children [5]. In Indonesia, one of the districts that has a high maternal mortality rate is Bojonegoro Regency which ranked third in the national scale in 2016. Therefore, reducing to zero value of maternal mortality is one of the government's objectives, especially Bojonegoro Regency. There are several factors that can affect maternal mortality, including the percentage of coverage of K1 (maternal visit 1), the percentage of coverage of K4 (visit 4), the percentage of obstetric complications, and the percentage of postpartum maternal services. We aim at examining the relationship of zero excess and overdispersion assumption violation and using simulation to determine which condition that is usually best for data with overdispersion.

## 2. The Modeling

There are explanation about ZINB and its analysis in details in the following sub chapters.

### 2.1. Zero Inflated Negative Binomial

There are two states in the ZINB regression, namely the emergence of zero values in the response variable [8]. First is zero state with probability of $\pi_i$. The second state is a negative binomial state where the response variable has a value following the negative binomial distribution with an average of $\mu$ with the chance of occurrence $(1-\pi_i)$. The probability function of the ZINB model with k as many predictor variables as in equation (2.1).

$$P(Y_i = y_i) = \begin{cases} \pi_i + (1-\pi_i)\left(\frac{1}{1+\tau\mu_i}\right)^{\frac{1}{\tau}}, \text{for } y_i = 0 \\ (1-\pi_i)\frac{\Gamma\left(y_i+\frac{1}{\tau}\right)}{\Gamma\left(\frac{1}{\tau}\right)y_i!}\left(\frac{1}{1+\tau\pi_{i_i}}\right)^{\frac{1}{k}}\left(\frac{\tau\mu_i}{1+\tau\mu_i}\right)^{y_i}, \text{for } y_i > 0 \end{cases} \tag{2.1}$$

where $0 \le \pi_i \le 1, \mu_i \ge 0$, $\tau$ : dispersion coefficient with $k > 0$, and $\Gamma(.)$ : Gamma function. When $p_i = 0$, the response variable follows negative binomial with mean $\mu_i$ and dispersion coefficient $\tau$, such that the response variable is defined as $Y_i \sim NB\ (\mu_i, \tau)$. The value of $\mu_i$ $(i = 1,2, \dots, n)$ is defined as the equation 2.2.

$$\begin{aligned} \mu_i &= \exp(\boldsymbol{x_i}^T\boldsymbol{\beta}) \\ \pi_i &= \frac{\exp(x_i{}^T\gamma)}{1+\exp(x_i{}^T\gamma)} \\ (1-\pi_i) &= \frac{1}{1+\exp(\boldsymbol{x_i}^T\boldsymbol{\gamma})} \end{aligned} \qquad (2.2)$$

Hence, the ZINB regression model is obtained as written in the equation (2.3). The model for discrete data that follows the negative binomial distribution is in the first part of equation (2.3) and the model for zero excess is in the second part of the equation (2.3) below.

$$\begin{aligned} &\ln \widehat{\mu_\iota} = \widehat{\beta}_0 + \textstyle\sum_{j=1}^{p} \widehat{\beta}_j x_{ij},\ i = 1,2,\dots,n, \text{ and } j = 1,2,\dots,k \\ &logit\ \widehat{\pi_\iota} = \widehat{\gamma}_0 + \textstyle\sum_{j=1}^{p} \widehat{\gamma}_j x_{ij},\ i = 1,2,\dots,n, \text{ and } j = 1,2,\dots,k \end{aligned} \qquad (2.3)$$

where

$k$ : Number of predictor variables

$n$: Sample size

$\beta$: Parameter of *ZINB* regression, *negative binomial state*

$\gamma$: Parameter of *ZINB* regression, *zero state*

From equation (2.3) it can be seen that β and γ are parameters of the ZINB model. In order to find the value of the ZINB parameter, Maximum Likelihood Estimation (MLE) method was used. The likelihood function as in equation (2.4)

$$L(\gamma,\beta) = \begin{cases} \prod_{i=1}^{n} \frac{e^{x_i^T\gamma}}{1+e^{x_i^T\gamma}} + \frac{1}{1+e^{x_i^T\gamma}}\left(\frac{1}{1+\tau e^{x_i^T\beta}}\right),\ untuk\ y_1 = 0 \\ \prod_{i=1}^{n} \frac{1}{1+e^{x_i^T\gamma}} \frac{\Gamma\left(y_i+\frac{1}{\tau}\right)}{\Gamma\left(\frac{1}{\tau}\right)y_i!}\left(\frac{1}{1+\tau e^{x_i^T\beta}}\right)^{\frac{1}{\tau}}\left(\frac{\tau e^{x_i^T\gamma}}{1+\tau e^{x_i^T\gamma}}\right)^{y_i},untuk\ y_i > 0 \end{cases} \qquad (2.4)$$

Then the *lnlikelihood* function can be obtained as the following:

$$\ln L(k,\beta,\gamma) = \begin{cases} \sum\limits_{\boldsymbol{i=1}}^{\boldsymbol{n}} ln(\frac{e^{(x_i{}^T\gamma)}}{1+e^{(x_i{}^T\gamma)}} + \frac{1}{e^{(x_i{}^T\gamma)}}\left(\frac{1}{1+\tau e^{(x_i{}^T\boldsymbol{\beta})}}\right)^{\frac{1}{\tau}},untuk\ \ y_i = 0) \\ \sum\limits_{\boldsymbol{i=1}}^{\boldsymbol{n}} \ln(\frac{1}{1+e^{(x_i{}^T\gamma)}} \frac{\Gamma\left(y_i+\frac{1}{\tau}\right)}{\Gamma\left(\frac{1}{\tau}\right)\Gamma(y_i+1)}\left(\frac{1}{1+\tau e^{(x_i{}^T\boldsymbol{\beta})}}\right)^{\frac{1}{\tau}}\left(\frac{\tau e^{(x_i{}^T\boldsymbol{\beta})}}{1+\tau e^{(x_i{}^T\boldsymbol{\beta})}}\right)^{y_i}. \\ \qquad untuk\ y_i > 0) \end{cases}$$

Next, ln-likelihood function will be maximized with the method of expectation maximization (EM). The EM method consists of two stages, namely expectations and maximization using Newton-Raphson iterations.

### 2.2. Bootstrapping

Bootstraping is defined as a method that can work without requiring distribution assumptions because the original sample is used as a population [6]. Bootstrap can be used to overcome various problems in data such as small data, data that violates assumptions, or data that has no assumptions. The Bootstrap method is done by taking samples from the original sample with sample size the same as the original one which then done the resampling. The bootstrap resampling procedure is as follows:

1. Establishing an empirical distribution ($\widehat{F_n}$) of the sample with the probability of each $x_i$ is $\frac{1}{n}$ where $i = 1,2,\dots,n$.
2. Taking a Bootstrap random sample of size *n* from empirical distribution in step 1. The taken sample is named the first Bootstrap sample $X^{*1}$.
3. Calculating the statistics $\widehat{\theta}$ , desired from the Bootstrap $X^{*1}$, called $\theta_1^*$
4. Repeat steps 2 and 3 to B times so that we get $\theta_1^*, \theta_2^*, \dots, \theta_B^*$.

Repetition of the bootstrap method will produce different results. If it can be done using all possible samples, $n^n$, the result will be the same.

## 3. Data and Methodology

### *3.1. Data*

In this study two data sources will be used. First, secondary data will be used regarding the Maternal Mortality Rate based on sub-districts in Bojonegoro Regency in 2016 [4] which have been proven to be overdispersed due to excess zero. There are several factors that can affect maternal mortality, including the percentage of coverage of K1 (maternal visit 1), the percentage of coverage of K4 (visit 4), the percentage of obstetric complications, and the percentage of postpartum maternal services.

The second data is generated using the statistics obtained from analysis of the first data. The data will be generated with various data characteristics to review the ZINB regression capability in overcoming the problem of overdispersion. Characteristics that will be used to generate response variables include:

1. The average or lambda (λ) = (0.2, 0.4, 0.6, 0.8, 1.0, 5.0).
2. The probability of zeros in the response ($p$) = (0.4, 0.6, 0.8).
3. Sample size ($n$) = 200, 500 and 800

Predictor variable in the data is predictor variable from data with resampling. The resampling method used is the bootstrap method. The analysis of each condition will be repeated 100 times

### *3.2. Analysis Method*

The steps in analyzing the data for simulation are as the following:

1. Generating response variable data with lambda ($\lambda$ = (0.2, 0.4, 0.6, 0.8, 1.0, 5.0)), probability of zeros ($p$ = (0.4, 0.6, 0.8)), and sample size $n$ of 200, 500 and 800, with each setting is repeated 100 times.
2. Obtain predictor variables with sample size (n) of 200, 500 and 800, using the bootstrap resampling method.
3. Checking the fulfillment of the assumption that there is no multicollinearity in the predictor variable by looking at the value of VIF.
4. Modeling Poisson Regression
5. Examining the fulfillment of equidispersion assumptions in the data through the Pearson Chi-Square test.
6. Checking the proportion of the zero value in the response variable of each model that is generated in step 1. If the proportion of the value of zero exceeds 0.5 then it can be continued with ZINB regression analysis.
7. Estimating the ZINB regression parameter using the MLE method
8. Simultaneous test
9. Checking the goodness of fit of the models
10. Choosing the best model based on the criteria of smallest AIC
11. Interpreting the results of the estimation

## 4. Data Analysis and Result

The data analysis and result part consist of two sub sections: application of ZINB to the data of maternal mortality rate and the second one is simulation data analysis.

### *4.1. Application to the Data of Maternal Mortality Rate*

The results of the parameter estimates from the analysis of the data of maternal mortality rate, done using RStudio, are as the following:

$$\ln \hat{\mu} = \widehat{\beta_0} + \widehat{\beta_1}X_1 + \widehat{\beta_2}X_2 + \widehat{\beta_3}X_3 + \widehat{\beta_4}X_4 \tag{4.1}$$

$$\ln\hat{\mu} = 17{,}9343 + 0{,}0065X_1 - 0{,}0620X_2 - 0{,}0303X_3 - 0{,}1092X_4,$$

$$\operatorname{logit}\hat{\pi} = \widehat{\gamma_0} + \widehat{\gamma_1}X_1 + \widehat{\gamma_2}X_2 + \widehat{\gamma_3}X_3 + \widehat{\gamma_4}X_4$$
$$\operatorname{logit}\hat{\pi} = -11{,}677 - 18{,}206X_1 + 7{,}206X_2 + 3{,}513X_3 + 7{,}675X_4. \quad (4.2$$

From equations (4.1) and (4.2) it can be seen if every 1% increase in K1 Coverage (visit 1) can increase maternal mortality by equal to $e^{0.0065}$ or 1.0065, while for each 1% increase in K4 coverage (visit 4), obstetric complications that were handled, and postpartum maternal services could reduce the average maternal mortality rate. In addition, every 1% increase in coverage of K1 (visit 1) can cause a tendency to increase the number of maternal deaths compared to before. A 1% increase in K4 coverage (visit 4), obstetric complications handled, and postpartum maternal care can cause a tendency to decrease the number of maternal deaths compared to before.

### *4.2. Simulation Results*

The parameter $\beta_0$ describes the average condition of the data. The estimation results of the $\beta_0$ parameter can be seen in Figure 1.

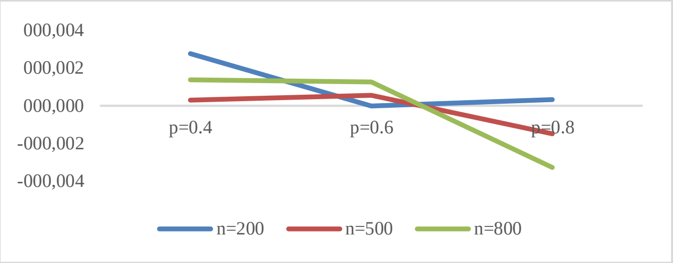
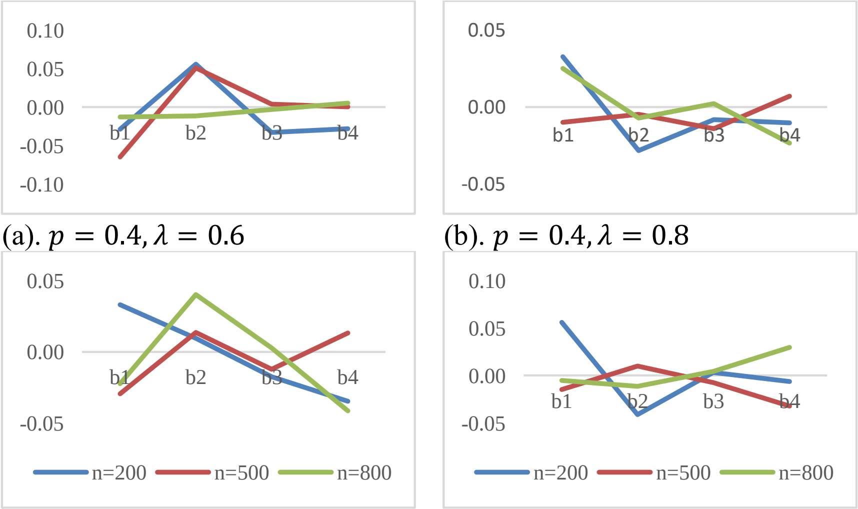


**Figure 1**. Average of Parameter Estimation Results $\beta_0$ with $\lambda = 0.6$.

From Figure 1 it can be seen if the increase in the value of $p$ is not accompanied by an increase in the results of parameter estimates or vice versa. However, the relatively large number of observations, namely $n$ with a value of 500 and 800 have a relatively similar pattern of estimation results compared to the results of parameter estimation with small observations. The results of estimating the ZINB Regression parameters for $\beta_1$, $\beta_2$, $\beta_3$, and $\beta_4$ can be seen in Figure 2.



(a). $p = 0.4, \lambda = 0.6$

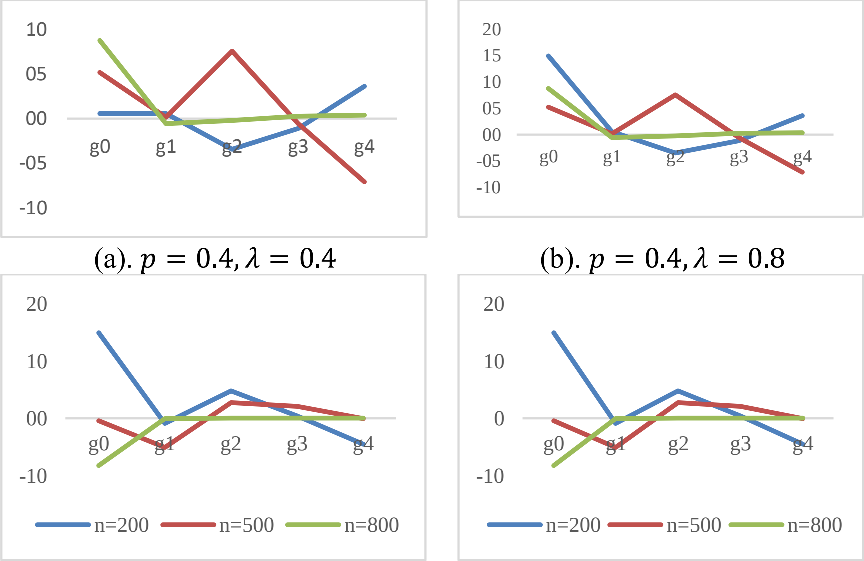
(b). $p = 0.4, \lambda = 0.8$

(c). $p = 0.6, \lambda = 0.6$

(d). $p = 0.6, \lambda = 0.8$

**Figure 2**. Estimates of $\beta_1$, $\beta_2$, $\beta_3$, and $\beta_4$.

From Figure 2, it can be seen if the estimation results of the parameters $\beta_1, \beta_2$, $\beta_3$, and $\beta_4$ are relatively more stable at n values of 500 and 800, which range from -0.02 to 0.02. However, this number is much smaller than the estimation of excess zero parameters. So if combined, it will be seen if the results of the discrete data parameter estimation are relatively the same for all generation conditions. The uniqueness of ZINB Regression is that it can accommodate excess zero conditions by performing modelling of $logit\ \hat{\pi}$ with regression parameters $\gamma_{0,}\ \gamma_1, \gamma_2, \gamma_{3,}$ and $\gamma_4$.



(a). $p = 0.4, \lambda = 0.4$ (b). $p = 0.4, \lambda = 0.8$

(c). $p = 0.6, \lambda = 0.4$ (d). $p = 0.6, \lambda = 0.8$

**Figure 3.** Estimates of $\gamma_{0,}\ \gamma_1, \gamma_2, \gamma_{3,}$ and $\gamma_4$.

From Figure 3, the results are in line with those obtained from the results of the discrete data parameter estimation where the greater the value of *n*, the value of parameter estimation results $\gamma_1, \gamma_2, \gamma_{3,}$ and $\gamma_4$ are in the same value range. But the value of the estimation result parameter $\gamma_0$ is very volatile. There were 54 conditions formed, but because they could not fulfil the excess zero requirement, three conditions could not be used in this study so there were 51 conditions used. Each condition was carried out 100 times. The selection of the best model is based on the AIC. With the help of the RStudio program package the results are shown in Tables 1 and 2.

**Table 1**. Average of *AIC*

| p | $\lambda$ | n | | |
|---|---|---|---|---|
| | | 200 | 500 | 800 |
| 0.6 | 0.2 | 128.4911 | 295.5576 | 469.9455 |
| | 0.4 | 191.2735 | 469.8743 | 749.6644 |
| | 0.6 | 248.6132 | 612.8341 | 990.0388 |
| | 0.8 | 303.5824 | 733.1905 | 1175.5410 |
| | 1.0 | 342.1938 | 838.2867 | 1333.0220 |
| | 5.0 | 625.7367 | 1552.1210 | 2473.4830 |

**Table 2**. Average of *AIC* (*n=200,* $\lambda = 0.2$)

| *p* | *AIC* |
|---|---|
| 0.4 | 164.3620 |
| 0.6 | 128.4911 |
| 0.8 | 80.8818 |

From Table 1 and 2 it can be seen that the greater the value of λ and *n*, the higher the average AIC value. The greater the *p* value will also result in a decrease in the average value of the AIC. The lowest AIC value is 80.8818 with a value of $n = 200$, $\lambda = 0.2$, and $p = 0.8$

## 5. Conclusion

From the result of the analysis, it can be concluded that an event of ovedispersion is always accompanied by the event of zero excess, but not vice versa. From the simulation process, several things can be obtained that the larger the value of $\lambda$, $p$, and $n$ then the dispersion coefficient gets larger and then overdispersion caused by excess zero can be well modeled with the ZINB regression as evidenced by the average value of $\tau$ less than 1 in all generation conditions. The best model is indicated by the lowest AIC value that is equal to 80.8818 with a combination of $n = 200$, $\lambda = 0.2$, and $p = 0.8$.

## References

[1] Draper, N. and Smith, H. 1992. *Analisis Regresi Terapan,* Translated to Indonesian. Jakarta : Gramedia.

[2] Rahayu, L. 2014. *Kajian Overdispersi Pada Regresi Poisson dan Zero Inflated Poisson untuk Beberapa Karakteristk Data.* Bogor : Sekolah Pasca Sarjana Institut Pertanian Bogor.

[3] Dewanti, N. 2016. *Perbandingan Zero Inflated Poisson (ZIP) dan Regresi Zero Inflated Negative Binomial (ZINB) pada Data Overdispersion (Studi Kasus : Angka Kematian Ibu di Provinsi Bali).* Denpasar : Fakultas MIPA Universitas Udayana.

[4] Prianggada, Q. 2017. *Pemilihan Regresi Zer-Inflated Poisson (ZIP) dan Zero Inflated Binomial Negative (ZIBN) Pada Angka Kematian Ibu Kabupaten Bojonegoro Tahun 2016 (Data Mengalami Overdispersi).* Malang: Fakultas MIPA, Universitas Brawijaya.

[5] Infodatin. 2014. *Situasi Kesehatan Ibu.* Jakarta : Pusat Data dan Informasi Departemen Kesehatan Republik Indonesia.

[6] Sungkono, J. 2013. *Resampling Bootstrap pada R. Thesis.* Klaten : Program Studi Pendidikan Matematikan FKIP UNWIDHA.

[7] Dinas Kesehatan. 2016. *Profil Kesehatan Kabupaten Bojonegoro tahun 2016.* Bojonegoro : Dinas Kesehatan Kabupaten Bojonegoro.

[8] Garray, A. M., Hashimoto, E. M., Ortega, E. M. M., dan Lachos, V. H. 2011. On Estimation and Influence Diagnostics For Zero Inflated Negative Binomial Regression Models. *Computational Statistics and Data Analysis*, **55** (3), p.1304-1318.