Media Statistika

p-ISSN 1979 – 3693 e-ISSN 2477 – 0647

MEDIA STATISTIKA 16(2) 2023: 206-214

http://ejournal.undip.ac.id/index.php/media_statistika

HANDLING OF OVERDISPERSION CASES IN MORBIDITY DATA IN SELUMA REGENCY

Mey Yanti Sarumpaet, Sigit Nugroho, Ramya Rachmawati

Statistics Study Program, The University of Bengkulu, Bengkulu, Indonesia

e-mail: meyyantisarumpaet.86@gmail.com

DOI: 10.14710/medstat.16.2.206-214

Article Info:
Received: 3 July 2023
Accepted: 20 February 2024
Available Online: 30 June 2024

Keywords:
Overdispersion, Poisson, Negative Binomial, Zero Inflated, Hurdle, Morbidity

Abstract: The problem of overdispersion as a violation of the assumption of equidispersion in Poisson Regression is generally caused by sources of unobserved heterogeneity, missing observations on predictor variables, outliers in the data, errors in the specification of the bridging function and many observed values that are zero. The purpose of this study is to find out the right model and the variables that affect data that occurs overdispersion and excess zero in the case of the number of days of disruption at work, school or other daily activities due to health complaints. The method used were Poisson Regression, Negative Binomial Regression, Hurdle Poisson Regression, Zero Inflated Poisson Regression, Zero Inflated Negative Binomial Regression and Hurdle Negative Binomial Regression. The data used were morbidity taken from data on the number of days of disruption at work, school or other daily activities due to health complaints in Seluma district, Bengkulu Province. It was found that the best model is Zero Inflated Negative Poisson with the smallest Akaike Information Criterion (AIC) value of 1620.609 and the variables that have a significant effect on the log model and the logit model are marital status and work variables.

1. INTRODUCTION

Poisson Regression Analysis is a method that can be used if the response variable is in the form of count data that follows a Poisson distribution, that is, when the data has a mean value that is roughly the same as the variance value (equidispersion). However, some research data found that the average value was greater than the variant value (overdispersion) (Fitriani et al., 2019). If Poisson regression is used in overdispersion conditions, it will result in errors in drawing conclusions, namely the explanatory variable looks as if it affects the response variable but actually has no effect (underestimates). Therefor Poisson regression is not appropriate for overdispersion data.

Several regression models used to overcome the overdispersion problem are: Hurdle Poisson (HP) regression model, Negative Binomial Regression model, Zero Inflated Poisson (ZIP) Regression model, Zero Inflated Negative Binomial (ZINB) Regression model and Hurdle Negative Binomial (HNB) model. Eunice et al. (2017) compared the Poisson regression model with Negative Binomial regression in a case study of malaria incidence in the village of Apac, Uganda and the results of this study showed that the Negative Binomial regression was better than the Poisson regression. According to Sreelatha & Miniswamy (2018), Zero Inflated Poisson (ZIP) regression is better than Poisson regression for overdispersion conditions, with many observations having a zero value on the response variable. Another study by Ariawan et al. (2012) used ZINB regression to overcome overdispersion due to excess zero in PT Sinar Mas Insurance Semarang Branch in 2010.

Another model used to overcome the overdispersion problem with an excess of zeros is the Hurdle Negative Binomial (HNB) model. This HNB model can overcome excess zero events by dividing the two models into two parts, namely the first model is a binary model for observations where the data is zero, namely the logit model. The second model for positive value data will be estimated using the Truncated Poisson model. This model has been used by Desjardins (2014), with a comparison of the performance of the ZINB regression model and the HNB regression using simulated data generated based on sample size, dispersion parameter values, parameter values of the first component (δ), the second component was conducted by Fitriani et al. (2019) count data with zero excess and overdispersion can be used in the Zero Inflated Negative Binomial (ZINB) regression model and the Hurdle Negative Binomial (HNB) regression model.

Research by Hu et al. (2011) on applications in the health sector experienced overdispersion and excess zero. The results show that ZINB and HNB have the best accuracy compared to the Poisson Regression, Negative Binomial (NB), ZIP, and Hurdle Poisson (HP) models. Yang et al. (2017) examined sick days in the State of Rhode Island, United States of America, comparing the ZIP, ZINB, HP, and HNB regression models. The results show that the ZINB and HNB regression models produce better performance than that of the other regression models.

The regression model that can overcome overdispersion is Negative Binomial regression, Zero Inflated Poisson (ZIP). Other models used to overcome the problem of overdispersion with excessive zero values are the Hurdle Negative Binomial (HNB), Zero Inflated Negative Binomial (ZINB), Hurdle Poisson (HP) models. The aim of this research is to compare the Hurdle Poisson (HP) regression model, the Negative Binomial Regression model, the Zero Inflated Poisson (ZIP) Regression model, the Zero Inflated Negative Binomial (ZINB) Regression model and the Hurdle Negative Binomial (HNB) model with AIC criteria (Akaike's Information Criterion) as the best model criteria. In addition, this study also aims to determine what variables affect the overdispersion and excess zero data.

Seluma Regency has the fifth highest percentage of residents who have health out of 10 districts in Bengkulu Province with a value of (24.19%) meaning that there are 24.19% of the population of Seluma Regency who experience health complaints so that daily activities or activities are disrupted.

2. LITERATURE REVIEW

2.1. Poisson Regression

Poisson regression is part of the regression analysis used to describe the relationship between the Y variable with the Poisson distribution and the independent variable X. The probability function of the Poisson distribution with the parameter μ where μ is the average of an event per unit time and t is a certain time period. The probability density function is

$$f_Y(y\,;\,\mu) = \begin{cases} \frac{e^{-\mu}\,\mu^y}{y!} & , \quad y = 0, 1, 2, \dots \\ 0 & , \qquad \text{other} \end{cases} \tag{1}$$

where μ is the average number of successes over a certain period of time or area.

2.2. Overdispersion

Overdispersion is a situation where the average value is greater than the variance value. According to Hardin and Hilbe (2008) overdispersion can occur due to data outliers, missing data observations in the explanatory variables, errors in determining the link function, correlations between observations, or transformations in the explanatory variables.

The test statistics used is

$$(\widehat{\Phi}) = \frac{\text{deviance}}{n-p} \tag{2}$$

where the deviance $= D = 2\sum_{i=1}^{n}\left(y_i log\frac{y_i}{\mu_i} - (y_i - \mu_i)\right)$

D is the deviation value, y_i is the value of the response variable from the i-th observation, μ_i is the average of the i-th cases in the Poisson Regression model, n is the number of observations and p is the number of parameters.

2.3. Excess zero

The problem in Poisson regression is excess zeros (Winkelmann, 2008). According to Winkelmann (2008) data with a large proportion of zero values when compared to other data values will result in precision in decision making.

Excess zero tests can be calculated using (Broek, 1995)

$$X_{cal}^2 = \frac{(n_0 - n\hat{p}_0)^2}{n\hat{p}_0(1-\hat{p}_0) - n\bar{x}{\hat{p}_0}^2} \tag{3}$$

with H_0 there is no excess zero in the data, H_1 there is excess zero in the data,

$\bar{x} = \hat{\lambda}$ is the average of count data, n_0 and n are the number of zero data and the number of observations consecutively. Note that $\hat{p}_0 = exp(-\hat{\lambda}\,)$. Reject H_0 whenever $X_{cal}^2 > X_{1;\alpha}^2$.

2.4. Negative Binomial Regression (NB Regression)

According to Hilbe (2011), NB regression is an applied model of GLM because NB regression is a family of exponential distributions. The response variable in the NB regression is assumed to follow the NB distribution. Let Y be the NB distribution, then the probability function is

$$P(Y) = \binom{y+k-1}{y} p^k (1-p)^y$$

$$= \frac{(y\,+\,k-1)}{(k-1)!\,y!} p^k (1-p)^y, \quad y = 0, 1, 2, \dots \tag{4}$$

where k is the number of successful events, y is the number of failed events before the k-th successful event occurs, p equals the probability of success and $(1-p)$ is the probability of failure.

According to McCullagh and Nelder (1989), an approach for overdispersion calculated data can use NB regression, because the NB regression is a mixed Poisson-

Gamma distribution. Let $y|\mu \sim Poisson\ (\mu)$ and $\mu \sim Gamma\ (\alpha\ ,\beta)$. Then the probability mass function of the mixed Poisson-Gamma distribution is

$$P(y|\alpha,\beta) = \int_0^\infty \text{Poisson}\ (y|\mu)\ .\text{Gamma}\ (\mu|\alpha\ ,\beta)\ d\mu$$

$$= \int_0^\infty \frac{e^{-\mu}\mu^{-y}}{y!} \cdot \frac{1}{\Gamma(\alpha)\beta^\alpha}\mu^{\alpha-1}exp\left(-\frac{\mu}{\beta}\right)d\mu$$

$$= \frac{\Gamma(y+\alpha)}{y!\,\Gamma(\alpha)}\left(\frac{1}{1+\beta}\right)^\alpha \left(1-\frac{1}{1+\beta}\right)^y \qquad (5)$$

where $\Gamma(\alpha)$ is the Gamma function of the number of successful occurrences. The parameter values of the Poisson-Gamma mixed distribution to form the NB regression model are expressed in the form $\mu = \alpha\beta$ and $k = 1/\alpha$, so that the mean and variance can be written as $E(Y) = \mu$ and $V(Y) = \mu + k\mu^2$, with k dispersion parameters, so the probability mass function of NB is

$$p(y,\mu,k) = \frac{\Gamma\left(y+\frac{1}{k}\right)}{\Gamma\left(\frac{1}{k}\right)y!}\left(\frac{1}{1+k\mu}\right)^{\frac{1}{k}}\left(\frac{k\mu}{1+k\mu}\right)^y, \qquad y = 0,1,2,\ldots \qquad (6)$$

As $k \to 0$, then $V(Y) \to \mu$ so that the NB distribution will approach Poisson regression where the variance and mean values are the same $E(Y) = V(Y) = \mu$, and if $k > 0$ the variance value will exceed the mean value or it is called overdispersion, $V(Y) > E(Y)$ (Greene, 2008).

2.5. Hurdle Poisson Regression (HPR)

HP regression is used to model count data with equidispersion or overdispersion conditions. The Hurdle model is able to overcome cases of excess zeros by dividing them into two models:

1. Binary data with a zero or positive value, this data will be interpreted using a logit model
2. Data with a positive value will be interpreted using truncated models

Suppose $k_1 = 0$ is the opportunity value when the response variable $(Y) = 0$ and $k_2(y)$, with $y = 1,2,\ldots$ is the opportunity function when the response variable (Y) is positive (Saffari et al., 2012), then the probability density function is

$$P(Y_i = y) = \begin{cases} k_1(0) & , \quad y = 0 \\ \left(1-k_1(0)\right)k_2(y) & , \quad y = 1,2,\ldots \end{cases} \qquad (7)$$

Suppose $Y_i,\ i = 1,2,\ldots n$ are response variables with non-negative count data and $Y_i = 0$ are response variables with excess zero values that the usual Poisson regression model cannot handle, so the distribution of Hurdle Poisson regression (Saffari, et al., 2012),

that is $P(Y_i = y_i) = \begin{cases} 1-\pi_i & , \quad y_i = 0 \\ (\pi_i)\frac{e^{-\mu_i}\mu_i{}^{y_i}}{(1-\pi_i)y_i!} & , \quad y_i > 0 \end{cases}$

where $0 < \pi_i < 1$ and $\pi_i = \pi_i(x_i)$, so that logit model with j-th variable as a response variable, that is $\pi_i = \frac{\exp\left(\sum_{j=1}^{p} z_{ij}\delta_j\right)}{1+\exp\left(\sum_{j=1}^{p} z_{ij}\delta_j\right)}$. Truncated Poisson models can be written, $\log(\mu_i) = \sum_{j=1}^{p} x_{ij}\beta_j$ or $\mu_i = \exp\left(\sum_{j=1}^{p} x_{ij}\beta_j\right)$.

2.6. Zero-Inflated Poisson Regression (ZIPR)

Jansakul & Hinde (2002) stated that if Y_i is an independent random variable with a ZIP distribution, then the zero value in the observation is assumed to appear in two ways that are appropriate for separate states. Taufan et al. (2012) stated that the parameter estimation of the Zero-Inflated Poisson regression uses the Maximum Likelihood Estimation (MLE) method, and the ln likelihood function equation is

$$P(Y_i = y_i) = \begin{cases} \sum_{i=1}^{n} \ln\left(e^{X_i^T\gamma} + \exp\left(-e^{X_i^T\beta}\right)\right) - \sum_{i=1}^{n} \ln\left(1 + e^{X_i^T\gamma}\right) & , \; y_i = 0 \\ \sum_{i=1}^{n} (X_i^T\beta) y_i - e^{X_i^T\beta} - \sum_{i=1}^{n} \ln\left(1 + e^{X_i^T\gamma}\right) - \sum_{i=1}^{n} \ln y_i\,! & , \; y_i > 0 \end{cases} \quad (8)$$

2.7. Zero Inflated Negative Binomial Regression (ZINBR)

Zero Inflated Negative Binomial (ZINB) regression is a model formed from a mixed Poisson and Gamma distribution, with its probability density function, Hilbe (2011), namely $f(y|\alpha,\beta) = \frac{\Gamma(y+\alpha)}{y!\Gamma(\alpha)}\left(\frac{1}{1+\beta}\right)^{\alpha}\left(1-\frac{1}{1+\beta}\right)^{y}$, $y = 0, 1, 2, \ldots$ with the mean and variance of the Negative Binomial distribution is $E[Y] = \alpha\beta$ and $\mathrm{V}[Y] = \alpha\beta + \alpha\beta^2$.

Garay & Hashimoto (2011) states that the first state of the ZINB regression is called the zero state with probability p_i and the observation results are zero, and the second state is called the negative binomial state with probability $(1 - p_i)$ and has a Negative Binomial distribution where the mean μ with $0 \le p_i \le 1$, so the probability density function is:

$$P(Y_i = y_i) = \begin{cases} p_i + \left(1 - p_i\right)\left(\frac{1}{1+k\mu_i}\right)^{1/k} & , \quad y_i = 0 \\ \left(1 - p_i\right)\frac{\Gamma\left(y+\frac{1}{k}\right)}{\Gamma\left(\frac{1}{k}\right)\Gamma\left(y_i+1\right)}\left(\frac{1}{1+k\mu_i}\right)^{1/k}\left(\frac{k\mu_i}{1+k\mu_i}\right)^{y_i} & , \quad y_i = 1, 2, 3, \ldots \end{cases} \quad (9)$$

with the assumption that μ_i and p_i depend on x_i and z_i variables, so that the model of ZINBR is divided into two:

1. Discrete data model for μ_i
   $\ln(\mu_i) = x_i^T\beta, \mu_i \ge 0, i = 1, \ldots n$. where x_i is a variable matrix that contains different sets of experimental factors related to the probability of a negative binomial mean in a negative binomial state.
2. Zero-Inflation Model for p_i
   $\text{logit}\ (p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = x_i^T\gamma, \qquad 0 \le p_i \le 1,\ i = 1, \ldots n$
   where x_i is a variable matrix that contains different sets of experimental factors associated with zero state probabilities.

2.8. Hurdle Negative Binomial Regression (HNBR)

The HNB regression model is used for the dependent variable in the form of count data and has more zero values than other values (excess zero) and experiences overdispersion (Desjardins, 2013). The HNB model uses a two-part approach (two-part model), namely the first part estimates a zero-value dependent variable called the Hurdle model and the second part estimates a non-negative round-valued dependent variable called the truncated model (Saffari, Adnan and Greene, 2012). Suppose y_i $(i = 1, 2, \ldots n)$ is a response variable in the form of count data $\left(y_i = 1, 2, \ldots n\right)$, then the probability function of the HNB regression model is

$$P(Y_i = y_i) = \begin{cases} \pi_i & , \quad y_i = 0 \\ \frac{(1-\pi_i)}{1-\left(\frac{k}{\mu+k}\right)^k} \frac{\Gamma(y+k)}{\Gamma(y+1)\Gamma(k)} \left(\frac{k}{\mu+k}\right)^k \left(1 - \frac{k}{\mu+k}\right)^y & , \quad y_i > 0 \end{cases} \qquad (9)$$

where π_i is the opportunity for the first state, namely the emergence of a zero state and the probability $(1 - \pi_i)$ for the second negative binomial state with $0 < \pi_i < 1$. μi is the average of the negative binomial distribution with $\mu_i > 0$ and k is the dispersion parameter that does not depend on the independent variable with $k > 0$.

The value of π_i and μ_i depends on the explanatory variables which can be defined as follows:

$$\pi_i = \left(\frac{e^{\boldsymbol{x}_i^T \boldsymbol{\delta}}}{1+e^{\boldsymbol{x}_i^T \boldsymbol{\delta}}}\right) \text{ and } \mu_i = e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}$$

so that the model for the connecting function logit and log is expressed as follows:

$$\text{logit } (\pi_i) = \log\left(\frac{\pi_i}{1 - \pi_i}\right) = \boldsymbol{x}_i^T \boldsymbol{\delta} \, ; \; i = 1, 2, .., n$$

logit $(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \boldsymbol{x}_i^T \boldsymbol{\delta} \, ; \; i = 1, 2, .., n$, so that $\pi_i = \frac{e^{\boldsymbol{x}_i^T \boldsymbol{\delta}}}{1+e^{\boldsymbol{x}_i^T \boldsymbol{\delta}}}$, for the value μ_i is $\log(\mu_i) = \boldsymbol{x}_i^T \boldsymbol{\beta}$, $\mu_i = e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}$.

For the *hurdle* model, the link function is logit $(\pi_i) = \hat{\delta}_0 + \sum_{j=1}^{p} x_{ij} \, \hat{\delta}_j$ where $i = 1, 2, \ldots n$ and $j = 1, 2, \ldots p$. The link function for Truncated Negative Binomial Model is $\log(\mu_i) = \hat{\beta}_0 + \sum_{i=1}^{p} x_{ij} \, \hat{\beta}_j$ where $i = 1, 2, \ldots n$ and $j = 1, 2, \ldots p$.

2.9. Morbidity

The concept of Statistics Indonesia in the Statistical Referral System (SIRUSA) Morbidity is defined as disturbances to physical or mental conditions, including accidents, or other things that disrupt daily activities. The morbidity rate has a more important role than the mortality rate, because if the morbidity rate is high it will result in death so that the mortality rate is high as well as the high morbidity value of an area indicates the poor health of the population in that region and vice versa. The lower the morbidity value indicates the health of the population in the region better.

3. RESEARCH METHODS

3.1. Data

The data used in this study is regarding morbidity obtained from the activities of the March 2021 National Socioeconomic Survey (Susenas), which is the number of days of disruption to health due to health complaints.

3.2. Variable

The variables used in this study were the number of days of health disturbance due to health complaints (Y) with the data from the March 2021 National Socio-Econimic

Survey (Susenas) Produced by the BPS of Seluma Regency, Bengkulu Province and the predictor variables are Gender (X_1), Education (X_2), Marital Status (X_3), Work (X_4).

3.3. Data Analysis

Data analysis uses R software (R Core Team, 2021), with the following steps: (1) Conduct descriptive analysis on research variables; (2) Overdispersion and under dispersion testing. If the data occurs overdispersion, then proceed to see whether the variable (Y) has an excess zero or not by looking at the proportion of its zero value; (3) Estimation of Poisson regression parameters, Negative Binomial regression, Hurdle Poisson regression, Zero Inflated Negative Binomial (ZINB), and Hurdle Negative Binomial (HNB) based on the smallest AIC value; (4) Identify variables that have a significant effect.

4. RESULT AND DISCUSSION

Seluma Regency is administratively included in the Bengkulu Province area. Seluma Regency is geographically located on the West Coast of Sumatra, which is at the latitude coordinates of 03°49´55.66"S - 04°1´40.22"S and longitude 101°17´27.57"E - 102° 59´40.54"E. Seluma Regency has an altitude between 0 until more than 1,000m above sea level. Seluma Regency is also included in the Bukit Barisan Hills which extends to the northwest - southeast with an altitude difference of about 300 m.

Excess zero testing is done by using the `zero.test` function in the `vcdExtra` package in the R program, with the null hypothesis H_0, namely there is no excess zero in the data and H_1, namely there is an excess zero. Based on the test results, the p-value is $2.22e^{-16}$ so that H_1 is accepted, which means that there is an excess zero in the data. Meanwhile, checking for overdispersion is done by dividing the deviation value by the degree of freedom. The result is that the dispersion value is 5.33 (> 1) meaning that the data has overdispersion. This indicates a violation of the assumptions that must be met when applying the Poisson regression model.

Table 1. AIC Comparison

Model	AIC
Poisson Regression	3025.113
Negative Binomial Regression (NBR)	1646.871
Hurdle Poisson Regression (HPR)	1929.574
Zero Inflated Poisson Regression (ZIPR)	1929.055
Zero Inflated Negative Binomial Regression (ZINBR)	**1620.609**
Hurdle Negative Binomial Regression (HNBR)	1621.434

Based on the smallest AIC criterion (Table 1), it is found that the best model to use for the data used and experiencing overdispersion and access zero is the Zero Inflated Negative Binomial Regression (ZINBR), with the equation

1. The discrete data model for $\widehat{\mu_i}$

$$\widehat{\mu_i} = \exp(-0.77806 + 0.32191\, X_1 + 0.08922X_2 + 1.52144X_3 - 0.56362X_4)$$

2. Zero Inflated Model (p_i) is

$$\hat{\mathrm{p}}_{\mathrm{i}} = \frac{\exp(-0.77806 + 0.32191\ \mathrm{X}_1 - 0.08922\ \ \mathrm{X}_2 + 1.52144\ \ \mathrm{X}_3 - 0.56362\ \ \mathrm{X}_4)}{1 + \exp(-0.77806 + 0.32191\ \mathrm{X}_1 - 0.08922\ \ \mathrm{X}_2 + 1.52144\ \ \mathrm{X}_3 - 0.56362\ \ \mathrm{X}_4)}$$

Based on the results of the equation then

1. Variables that significantly affect on the count model include the marital status variable (X_3), where people with marital status have a number of days of impaired health $\exp(1.52144) = 4.57881$ times longer than people who have not married. The next variable that has a significant influence is work (X_4). People who have worked have fewer days of health problems $\exp(-0.56362) = 0.56914$ than people who do not work. These estimates use other variables Gender (X_1) and Education (X_2) also in the model.
2. For the zero inflation data model (pi) the variable that has a significant effect is the marital status variable (X_3), with the interpretation that married people have fewer days of health problems $\exp(-3.28128) = 0.03758$ times less of unmarried people, meaning that the number of days of disruption to health for unmarried people is 26.60975 longer than for married people. The next variable that affects the number of days of health disturbance is work (X_4). People who have jobs have an influence of $\exp(2.10121) = 8.17604$ longer than people who do not have jobs. These estimates using other variables Gender (X_1) dan Education (X_2) also in the model.

5. CONCLUSION

From the comparison of the AIC values of the Poisson regression model, the Negative Binomial Regression, the ZINB Regression and the HNB Regression model, it shows that the AIC value of the ZINB regression model is the smallest, namely 1620,609. Based on this, it can be concluded from the four regression models used in this study, the ZINB Regression model is best used to model data on the number of days of disruption of daily activities due to health complaints in Seluma District which contain overdispersion and excess zero.

The variables that affect the number of days of disruption of daily activities due to health complaints in Seluma District in the ZINB Regression model with the Negative Binomial with log model (log model) are marital status and work status variables. Meanwhile, the ZINB regression uses the logit model using predictors of marital status and work variables.

REFERENCES

Ariawan, B., Suparti, & Sudarno. (2012). Pemodelan Regresi Zero-Inflated Negative Binomial (ZINB) untuk Data Respon Diskrit dengan Excess Zeros. *Jurnal Gaussian*, *1*(1), 55-64. https://doi.org/10.14710/j.gauss.1.1.55-64

Broek, J. (1995). A Score Test of for Zero Inflation In a Poisson Distribution. *Biometrics, 51*, 738-743. https://doi.org/10.2307/2532959

Desjardins, C. D. (2014). Evaluating the Performance of Two Competing Models of School Suspension Under Simulation - the Zero-inflated Negative Binomial and the Negative Binomial Hurdle. *Dissertation Abstracts International Section A: Humanities and Social Sciences*, *74*(10-A(E)).

Eunice, A., Wanjoya, A., & Luboobi, L. (2017). Statistical Modeling of Malaria Incidences in Apac District, Uganda. *Open Journal of Statistics*, *07*(06), 901-919. https://doi.org/10.4236/ojs.2017.76063

Fitriani, R., Chrisdiana, L. N., & Efendi, A. (2019). Simulation on the Zero Inflated Negative Binomial (ZINB) to Model Overdispersed, Poisson Distributed Data. *IoP Conference Series: Materials Science and Engineering*, *546*(5), 52025. https://doi.org/10.1088/1757-899X/546/5/052025

Garay, A., Hashimoto, E., Lachos, V., & Ortega, E. (2011). On Estimation and Influence Diagnostics for Zero-Inflated Negative Binomial Regression Models. *Computational Statistics and Data Analysis*, *55*, 1304-1318

Greene, W. (2008). Functional Forms for the Negative Binomial Model for Count Data. *Economics Letters*, *99*(3), 585-590. https://doi.org/10.1016/j.econlet.2007.10.015

Hardin, J. W., & Hilbe, J. M. (2008). Analysis of Fit. In *Generalized Linear Models and Extensions, Second Edition* (Vol. 2). Texas: Stata Press.

Hilbe, J. M. (2011). *Negative Binomial Regression* (2th ed.). New York: Cambridge University Press.

Hu, M. C., Pavlicova, M., & Nunes, E. V. (2011). Zero-inflated and Hurdle Models of Count Data with Extra Zeros: Examples from an HIV-risk Reduction Intervention Trial. *American Journal of Drug and Alcohol Abuse*, *37*(5), 367-375. https://doi.org/10.3109/00952990.2011.597280

Jansakul, N. & Hinde, J. P. (2002). Score Tests for Zero-Inflated Poisson Models. *Computational Statistics & Data Analysis*, *40*, 75-96.

Mc.Cullagh, P. & Nelder, J. A. (1989). *Generalized Linear Models* (2th ed). London: Chapman and Hall.

R Core Team (2021). R: A Language And Environment For Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Saffari, S. E., Adnan, R., & Greene, W. (2012). Parameter Estimation on Hurdle Poisson Regression Model with Censored Data. *Jurnal Teknologi (Sciences and Engineering)*, *57*(1),189-198. https://doi.org/10.11113/jt.v57.1533

Sreelatha, C. H., & Muniswamy, B. (2018). A Wald Test for Over Dispersion in Zero-Inflated Poisson Regression Model. In *International Journal of Mathematical Archive*, *9*(6), 201–212.

Taufan M., Suparti, & Agus R., (2012). Analisis Faktor-Faktor yang Mempengarhi Banyaknya Klaim Assuransi Kendaraan Bermotor Menggunakan Model Regresi *Zero-Inflated Poisson* (Studi Kasus di PT. Asuransi Sinar Mas Cabang Semarang Tahun 2010). *Media Statistika*, *5*(1), 49-61. https://doi.org/10.14710/medstat.5.1.49-62

Winkelman, R. (2008). *Econometric Analysis of Count Data*, *5th edition*. Berlin: Springer.

Yang, S., Puggioni, G., Harlow, L. L., & Redding, C. A. (2017). A Comparison of Different Methods of Zero-Inflated Data Analysis and an Application in Health Surveys. *Journal of Modern Applied Statistical Methods*, *16*(1), 518-543. https://doi.org/10.22237/jmasm/1493598600