Negative Binomial Regression Analysis on Dengue Hemorrhagic Fever Cases in East Kalimantan Province

M. Fathurahman[a)], Ika Purnamasari, and Surya Prangga

Department of Mathematics, Faculty of Mathematics and Natural Sciences, Mulawarman University, Samarinda, East Kalimantan, Indonesia.

[a)] Corresponding author: fathur@fmipa.unmul.ac.id

Abstract. Poisson regression is commonly used in modeling count data. An essential assumption of the Poisson regression model is that the mean of the response variable is equal to the variance, namely equidispersion. Many fields of research were the data overdispersed, which is the variance greater than its mean. Therefore, the Poisson regression model is not suitable to model it. The negative binomial regression (NBR) model is a solution of the Poisson regression model when the response variable is an overdispersion count data. This study aims to build an NBR model and apply it to model the dengue hemorrhagic fever (DHF) cases in East Kalimantan Province, Indonesia, in 2019. The maximum likelihood estimation and Fisher scoring methods were used to estimate the NBR model parameters, whereas the significant test containing the overall and individual tests was done using the likelihood ratio and Wald test statistics. Based on data analysis, the mean and variance values of the DHF data in East Kalimantan Province were 672 and 386113, respectively, and it shows that the DHF cases in East Kalimantan Province, Indonesia, in 2019 were an overdispersed count data. The factors that affected the DHF cases in East Kalimantan Province, Indonesia, in 2019 based on the NBR model were the total area, the area altitude, population density, and the health workers.

INTRODUCTION

Count data is one type of statistical data that shows the number of events over a particular time and can only be positive so that the conventional method can be used in the modeling with Poisson regression. Some assumptions must be met when using Poisson regression. The mean and variance must be equal, namely *equidispersion* [1]. In the case of count data, this assumption is often not met because many data in various research fields that are the variance is greater than the mean, called *overdispersion*. An invalid model can underestimate standard errors and misleading inference for regression parameters [2]. Therefore, an approach is needed to overcome the problem of overdispersion in Poisson regression. Several studies that can be used to accommodate overdispersion in Poisson regression have been proposed. The negative binomial regression (NBR) model was proposed [3,4,5]. The NBR model was constructed with a mixed Poisson and gamma distribution. Meanwhile, the random-effects regression models for handling overdispersion due to latent heterogeneity were founded [6,7].

Furthermore, the dengue hemorrhagic fever (DHF) cases sufferers is a count data that tends to be overdispersion. DHF is a major public health problem in Indonesia, and the DHF mortality rate is continually increasing from year to year. DHF often appears as an extraordinary event (KLB) with a relatively high mortality rate. The Aedes Aegypti mosquito vector transmits DHF through bites. The most common places for these mosquitoes are humid, high rainfall, puddles of water inside and outside the house. Another factor that causes dengue fever is population density and unhealthy community behavior [8].

According to [9], the DHF cases in Indonesia in 2019 were 138,127. This number increased compared to 2018 of 65,602 cases. Deaths due to DHF in 2019 also increased compared to 2018, which was 467 to 919 deaths. Illness and death can be described using the incidence rate (IR) indicator per 100,000 population and the case fatality rate (CFR) as a percentage. The IR of DHF in Indonesia in 2019 was 51.48 per 100,000 population. It describes an increase compared to the previous two years. Of all provinces in Indonesia, in 2019, East Kalimantan Province ranks

The 3rd International Conference on Mathematics and Sciences (The 3rd ICMSc)
AIP Conf. Proc. 2668, 070002-1–070002-6; https://doi.org/10.1063/5.0111999
Published by AIP Publishing. 978-0-7354-4214-6/$30.00

second for the highest IR, with an IR of 180.66. This IR value also exceeds Indonesian IR. It shows that the DHF cases in East Kalimantan Province are high compared to other provinces in 2019.

The purpose of this study is to examine the theory and application of the NBR model. A study of theory discusses the estimation and testing of parameter hypotheses. The maximum likelihood estimation (MLE) and Fisher scoring methods estimate the NBR model parameters. Hypothesis testing using the likelihood ratio test (LRT) and Wald test methods. The applied study is modeling the factors that influence the DHF cases in East Kalimantan Province, Indonesia, in 2019.

MATERIALS AND METHODS

Data Sources and Research Variables

The data used in this study is secondary data obtained from the Central Bureau of Statistics [10]. The research variables used in this study are presented in Table 1.

TABLE 1. Research variables

Symbols	Variables	Variable Types
Y	The DHF cases	Discrete
X_1	The total area	Numeric
X_2	The area altitude	Numeric
X_3	Population density	Numeric
X_4	The health workers	Discrete

Poisson Regression

Poisson regression is included in the generalized linear models [11]. The Poisson regression model can be used to model events that have a small probability of occurrence with occurrence depending on a certain time interval [12]. The response of the Poisson regression model is the count data which is assumed to follow the Poisson distribution with the probability mass function defined as follows [1]:

$$P(y;\mu) = \frac{e^{-\mu}\mu^y}{y!}, y = 0,1,2,\dots;\ \mu > 0 \tag{1}$$

where μ is the mean of the number of events in a certain interval. The mean and variance of the Poisson distribution based on Equation (1) are expressed by $E(Y) = Var(Y) = \mu$.

According to [11], the relationship function for Poisson regression is stated as follows:

$$\eta_i = \log(\mu_i) = \boldsymbol{x}_i^T\boldsymbol{\theta}, i = 1,2,\dots,n \tag{2}$$

where η_i is the link function at i-th observation, μ_i is the mean of response variable at i-th observation, $\boldsymbol{x}_i^T = [1 \quad x_{1i} \quad x_{2i} \quad \cdots \quad x_{ki}]$ is the vector of the explanatory variables at i-th observation, $\boldsymbol{\theta} = [\theta_0 \quad \theta_1 \quad \theta_2 \quad \cdots \quad \theta_k]^T$ is the vector of parameters, and n is the sample size.

Based on Equation (2), the form of the Poisson regression model is

$$y_i = \mu_i + \varepsilon_i = \exp(\boldsymbol{x}_i^T\boldsymbol{\theta}) + \varepsilon_i, i = 1,2,\dots,n \tag{3}$$

where μ_i, $\boldsymbol{x}_i^T$, $\boldsymbol{\theta}$, and n as in Equation (2), y_i is the response variable at i-th observation, and ε_i is the error at i-th observation.

Overdispersion

[11] state that the count data contains overdispersion if the variance is greater than the mean, namely $Var(Y) > E(Y)$. Overdispersion occurs due to unobserved sources of variability in the data or the influence of other variables that result in the probability of an event occurring depending on previous events. Overdispersion can lead to underestimating the standard error, resulting in underestimated parameters and the significance of the explanatory variable effect being overestimated. Overdispersion in Poisson regression can be detected by the deviance divided by the degrees of freedom. If the value is greater than one, it is said that there is overdispersion in the data [1].

Negative Binomial Regression

Negative binomial regression (NBR) is one solution to overcome the overdispersion problem based on the Poisson-gamma mixture model [1, 13]. According to [14], the negative binomial distribution that accommodates overdispersion has a density function as follows:

$$P\big(y_{i;}\mu_i,\zeta\big)=\frac{\Gamma(y_i+\zeta^{-1})}{\Gamma(\zeta^{-1})\Gamma(y_i+1)}\left(\frac{\zeta^{-1}}{\zeta^{-1}+\mu_i}\right)^{\zeta^{-1}}\left(\frac{\mu_i}{\zeta^{-1}+\mu_i}\right)^{y_i},i=1,2,\dots,n;\ y_i=0,1,2,\dots;\ \zeta\geq 0, \tag{4}$$

where $\Gamma(\cdot)$ is the gamma function, and ζ is the dispersion parameter. The mean and variance of the negative binomial distribution are expressed by $E(y_i|x_i)=\mu_i$ and $Var(y_i|x_i)=\mu_i+\zeta\mu_i^2$, respectively.

The NBR model can be written as

$$\eta_i=\log(\mu_i)=\log[\exp(\boldsymbol{x}_i^T\boldsymbol{\theta})]=\theta_0+\theta_1x_{1i}+\theta_2x_{2i}+\cdots+\theta_kx_{ki}=\boldsymbol{x}_i^T\boldsymbol{\theta},i=1,2,\dots,n \tag{5}$$

where η_i, μ_i, $\boldsymbol{x}_i^T$, $\boldsymbol{\theta}$, and n as in Equation (2).

The NBR model can be obtained by estimating the model parameters using the MLE method. The initial step of parameter estimation using the MLE method is to form the likelihood and log-likelihood functions as follows:

$$\mathcal{L}(\boldsymbol{\theta},\zeta)=\prod_{i=1}^{n}\left[\frac{\Gamma(y_i+1/\zeta)}{\Gamma(1/\zeta)\Gamma(y_i+1)}\left(\frac{1}{1+\zeta\mu_i}\right)^{1/\zeta}\left(\frac{\zeta\mu_i}{1+\zeta\mu_i}\right)^{y_i}\right]. \tag{6}$$

Since $\frac{\Gamma(y_i+1/\zeta)}{\Gamma(1/\zeta)\Gamma(y_i+1)}=\prod_{j=1}^{y_i-1}(j+1/\zeta)$, the likelihood function in Equation (6) can be written as

$$\mathcal{L}(\boldsymbol{\theta},\zeta)=\prod_{i=1}^{n}\left[\left(\prod_{j=1}^{y_i-1}j/\zeta\right)\frac{1}{(y_i!)}\left(\frac{1}{1+\zeta\mu_i}\right)^{1/\zeta}\left(\frac{\zeta\mu_i}{1+\zeta\mu_i}\right)^{y_i}\right]. \tag{7}$$

$$\ell(\boldsymbol{\theta},\zeta)=\log[\mathcal{L}(\boldsymbol{\theta},\zeta)]=\sum_{i=1}^{n}\left[\left(\sum_{j=1}^{y_i-1}\log(j+1/\zeta)\right)-\log(y_i!)+y_i\log(\zeta\mu_i)-(y_i+1/\zeta)\log(1+\zeta\mu_i)\right]. \tag{8}$$

Furthermore, maximizing the log-likelihood function in Equation (8) by determining the first-order partial derivatives of the log-likelihood function, then equating them to zero:

$$\begin{aligned}\frac{\partial\ell(\boldsymbol{\theta},\zeta)}{\partial\theta_0}&=\sum_{i=1}^{n}\left[y_i-(y_i+1/\zeta)\left(\frac{\zeta\mu_i}{1+\zeta\mu_i}\right)\right]=\sum_{i=1}^{n}\left(\frac{y_i-\mu_i}{1+\zeta\mu_i}\right)=0.\\ \frac{\partial\ell(\boldsymbol{\theta},\zeta)}{\partial\theta_q}&=\sum_{i=1}^{n}\left[y_ix_{ip}-(y_i+1/\zeta)\left(\frac{\zeta\mu_ix_{ip}}{1+\zeta\mu_i}\right)\right]=\sum_{i=1}^{n}\left(\frac{\mu_i}{1+\zeta\mu_i}\frac{(y_i-\mu_i)x_{ip}}{\mu_i}\right)=0.\\ \frac{\partial\ell(\boldsymbol{\theta},\zeta)}{\partial\zeta}&=\sum_{i=1}^{n}\left[-\zeta^{-2}\sum_{j=0}^{y_i-1}\frac{1}{(j+\zeta^{-1})}+\zeta^{-2}\log(1+\zeta\mu_i)-\left(\frac{y_i-\mu_i}{\zeta(1+\zeta\mu_i)}\right)\right]=0.\end{aligned} \tag{9}$$

The second-order partial derivative of the log-likelihood function based on Equation (9) is as below.

$$\begin{aligned}\frac{\partial^2\ell(\boldsymbol{\theta},\zeta)}{\partial\theta_0^2}&=-\sum_{i=1}^{n}\left(\frac{(1+\zeta y_i)\mu_i}{(1+\zeta y_i)^2}\right)\\ \frac{\partial^2\ell(\boldsymbol{\theta},\zeta)}{\partial\theta_0\partial\theta_r}&=-\sum_{i=1}^{n}\left(\frac{(1+\zeta y_i)x_{ir}\mu_i}{(1+\zeta y_i)^2}\right),r\leq k\\ \frac{\partial^2\ell(\boldsymbol{\theta},\zeta)}{\partial\theta_l\partial\theta_r}&=-\sum_{i=1}^{n}\left(\frac{x_{il}x_{ir}\mu_i(1+\zeta y_i)}{(1+\zeta\mu_i)^2}\right),l\leq k\\ \frac{\partial^2\ell(\boldsymbol{\theta},\zeta)}{\partial\zeta^2}&=\sum_{i=1}^{n}\left[\zeta^{-3}\sum_{j=0}^{y_i-1}\frac{(2j+\zeta^{-1})}{(j+\zeta^{-1})^2}-2\zeta^{-3}\log(1+\zeta\mu_i)-\frac{\zeta^2\mu_i}{(1+\zeta\mu_i)}-\frac{(y_i-\mu_i)(1+2\zeta\mu_i)}{(\zeta+\zeta^2\mu_i)^2}\right],\end{aligned} \tag{10}$$

where k, l, and r are the number of parameters.

The maximum likelihood estimator of the NBR model parameters in Equation (9) has an implicit form. Therefore, we need the numerical approach. The Fisher scoring method was used [15]. The Fisher scoring algorithm for obtaining the maximum likelihood estimator is as follows:

1. Determine the initial value for $\hat{\boldsymbol{\theta}}$ and $\hat{\zeta}$, namely $\hat{\boldsymbol{\theta}}^{(0)} = \left[\theta_0^{(0)} \quad \theta_1^{(0)} \quad \theta_2^{(0)} \quad \cdots \quad \theta_k^{(0)}\right]^T$ and $\hat{\zeta}^{(0)}$.
2. Determine the tolerance value, symbolized by δ for the iteration process stopping.
3. Start the iteration process using the Fisher scoring formula:
$$\hat{\boldsymbol{\theta}}^{(u+1)} = \hat{\boldsymbol{\theta}}^{(u)} + \boldsymbol{I}^{-1}\left(\hat{\boldsymbol{\theta}}^{(u)}\right)\boldsymbol{g}\left(\hat{\boldsymbol{\theta}}^{(u)}\right), u = 0,1,2,\dots,m. \tag{11}$$
where $\boldsymbol{g}(\boldsymbol{\theta})$ is the gradient vector, which has the elements in Equations (9). $\boldsymbol{I}(\boldsymbol{\theta})$ is the information matrix and defined as
$$\boldsymbol{I}(\boldsymbol{\theta}) = E\left(-\frac{\partial^2 \ell(\boldsymbol{\theta},\zeta)}{\partial\theta_l \partial\theta_r}\right) = \sum_{i=1}^{n}\left(\frac{x_{il}x_{ir}\mu_i}{1+\zeta\mu_i}\right),$$
where the $\partial^2 \ell(\boldsymbol{\theta},\zeta)/\partial\theta_l\partial\theta_r$ displays in Equation (10).
4. The iteration stops at the m-th iteration if the condition of convergence is satisfied, which is $\left\|\hat{\boldsymbol{\theta}}^{(m+1)} - \hat{\boldsymbol{\theta}}^{(m)}\right\| \leq \delta$, where δ is the smallest positive value. The estimator values of the parameters are obtained in the last iteration.

Furthermore, hypothesis testing on the NBR model parameters was employed to get the explanatory variables influencing the response variable. Hypothesis testing contains an overall test and the individual test. The overall test is used to jointly obtain the significant effect of the explanatory variables on the response variable. Meanwhile, the individual test is used to get the individually significant effect of the explanatory variables on the response variable.

The test used for the overall test is the LRT method with the following hypothesis:
$H_0 : \theta_1 = \theta_2 = \cdots = \theta_k = 0$
$H_1 :$ at least one of $\theta_r \neq 0, r = 1,2,\dots,k.$
The test statistic used for this test is *Wilk's lambda* statistic which is defined as follows:
$$G_1 = 2\sum_{i=1}^{n}\left(y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) - \left(y_i + \frac{1}{\hat{\zeta}}\right)\log\left(\frac{1+\hat{\zeta}y_i}{1+\hat{\zeta}\hat{\mu}_i}\right)\right). \tag{12}$$

The test statistic in Equation (12) is an asymptotic chi-square distribution [16]. Therefore, the critical region of the test is the null hypothesis is rejected when Wilk's lambda statistic value is greater than the $\chi^2_{(\alpha,v)}$ value (i.e., $G_1 > \chi^2_{(\alpha,v)}$), where α is a significance level, and v is the degrees of freedom. The $\chi^2_{(\alpha,v)}$ value is obtained from the chi-square distribution table, which is $v = k - 2$. On the other hand, the null hypothesis is rejected when the p-value is less than α.

The following hypothesis test is an individual test using the Wald test procedure. The hypothesis is
$H_0 : \theta_r = 0$
$H_1 : \theta_r \neq 0, r = 1,2,\dots,k.$
The test statistic is Wald statistic, which is formulated by
$$G_2 = \frac{\hat{\theta}_r}{SE\left(\hat{\theta}_r\right)}, \tag{13}$$
where $SE\left(\hat{\theta}_r\right) = \sqrt{\widehat{Var}\left(\hat{\theta}_r\right)}$ is the standard error of $\hat{\theta}_r$. $\sqrt{\widehat{Var}\left(\hat{\theta}_r\right)}$ is the diagonal element of $\boldsymbol{I}^{-1}\left(\hat{\boldsymbol{\theta}}^{(u)}\right)$, and $\boldsymbol{I}(\boldsymbol{\theta})$ is derived in Equation (11). The Wald statistic in Equation (13) has an asymptotic standard normal distribution [16]. Thus, the null hypothesis is rejected when the Wald statistic value falls into the rejection region, namely, $|G_2| > Z_{\alpha/2}$. The $Z_{\alpha/2}$ value can be obtained from the table of standard normal distribution. If using the p-value, then the null hypothesis is rejected when the p-value is less than α.

RESULTS AND DISCUSSION

The descriptive statistics of the research variable includes the response variable and the explanatory variables, are given in Table 2.

TABLE 2. The descriptive statistics of the research variables

Variables	Minimum	Maximum	Mean	Standard Deviation
Y	66	1838	672	621
X_1	163.1	31051.7	12734.7	11494.05
X_2	5.98	174.63	54.1	65.04
X_3	1.36	1297.74	379.65	579.07
X_4	287	2960	1381	903

Based on Table 2, the highest DHF cases were in Balikpapan City, whereas the lowest was in Mahakam Ulu Regency. The largest area was in Kutai Timur Regency, and the smallest was Bontang City. The highest area was Mahakam Ulu Regency, and the smallest was Kutai Timur Regency. The most density of population was in Balikpapan City, and the sparsely was Mahakam Ulu Regency. Most health workers in Samarinda City and a few were in Mahakam Ulu Regency. Meanwhile, the mean and variance values of the DHF cases in Table 2 are not equal, which was the variance of the DHF cases greater than its mean. Consequently, the DHF cases in East Kalimantan Province were overdispersed count data, and the NBR can model it.

They were, furthermore, detecting the multicollinearity of the explanatory variables using the variance inflation factor (VIF) values. The VIF values of all explanatory variables in Table 3 are less than ten, which indicates no multicollinearity. Therefore, all explanatory variables can be used in the NBR model.

TABLE 3. VIF values of the explanatory variables

Explanatory Variables	VIF
X_1	2.8981
X_2	1.3499
X_3	5.1523
X_4	2.9927

Table 4 below presents the estimation and hypothesis testing results of the NBR model parameters used to model the factors influencing the DHF cases in East Kalimantan Province.

TABLE 4. Parameter estimates and the test statistic values of the overall test and the individual test

Parameter	Estimation	Standard Error	G_2	p-Value
θ_0	4.988	0.2849	17.541	< 0.0001
θ_1	2.757×10^{-5}	1.285×10^{-5}	2.145	0.03197
θ_2	-6.791×10^{-3}	1.587×10^{-3}	-4.278	< 0.0001
θ_3	9.609×10^{-4}	3.390×10^{-4}	2.835	0.00458
θ_4	5.240×10^{-4}	1.643×10^{-4}	3.190	0.00142

ζ = 15,6363
G_1 = 27.7541, p-value = 9.4031×10^{-7} (< 0.0001)
α = 0.05
v = 2
$\chi^2_{(\alpha,v)}$ = 5.9915
$Z_{\alpha/2}$ = 1.96

Wilk's lambda statistic (G_1) value of the overall test in Table 4 exceeded the $\chi^2_{(\alpha,v)}$ value. The p-value was less than α. The results indicated that the total area, the area altitude, population density, and the health workers were jointly significantly affecting the DHF cases in East Kalimantan Province. Meanwhile, the individual test was used to obtain the explanatory variables that significantly affect the DHF cases in East Kalimantan Province. Based on Table 4, all values of the Wald statistic exceeded the $Z_{\alpha/2}$ value, and also, all of the p-values were less than α. Therefore, the conclusion was that the total area, the area altitude, population density, and the health workers significantly affected the DHF cases in East Kalimantan Province.

CONCLUSION

Poisson regression is a popular choice in modeling count data. One of the essential assume in Poisson regression is equidispersion, where the mean of response is equal to the variance. Many data counts in various research fields are overdispersed, which is the response variable variance greater than its mean. Therefore, Poisson regression cannot be used to model it. Overdispersion response data can be modeled with NBR. The NBR model was obtained using the MLE and Fisher scoring methods. Hypothesis testing of the NBR model contains the overall test and the individual test. The overall test was employed by Wilk's lambda statistic, whereas the Wald statistic was used for the individual test. The NBR model was applied to the DHF cases in East Kalimantan Province, Indonesia, in 2019. The mean and variance of DHF cases have the values of 672 and 386,113, respectively, and it shows that the DHF cases were an overdispersion count data. Therefore, the NBR model was suitable to model. Based on the NBR model analysis, the factors significantly influencing the DHF cases in East Kalimantan Province, Indonesia, in 2019 were the total area, the area altitude, population density, and the health workers.

ACKNOWLEDGMENTS

We would like to thank Mulawarman University for funding the research through IsDB Grant Research 2021 under 405/UN17.11/PL/2021.

REFERENCES

1. J. M. Hilbe. Negative Binomial Regression, 2nd Ed. (Cambridge University Press, New York, 2011).
2. D. T. Molla and B. Muniswamy. *IOSR Journal of Mathematics.* 1, 29-36 (2012).
3. J. F. Lawless. *Canad. J. Statist.* 15, 209-225 (1987).
4. C. Dean and J. F. Lawless. *J. Amer. Statist. Assoc.* 84, 467-472 (1989).
5. P. Wang, M. L. Puterman, I. Cockburn, and N. Le. *J. biom.* 52, 381-400, (1996).
6. I. Ozmen. *Biom. J.* 42, 303-314 (2000).
7. Y. Lee, Y and J. A. Nelder. *J. Appl. Stat.* 49, 591-598 (2000).
8. Jumaina and A. Gani. "*Determinants of the Incidence of Dengue Hemorrhagic Fever in the Work Area of Kunciran Health Center, Tangerang, Banten*", in *The 6th International Conference on Public Health Proceedings*. (Masters Program in Public Health, Graduate School, Universitas Sebelas Maret, Surakarta, 2019).
9. Ministry of Health. Indonesian Health Profile in 2019. (Ministry of Health, Jakarta, 2020).
10. Central Bureau of Statistics. East Kalimantan Province in Figures 2020. (Central Bureau of Statistics of East Kalimantan Province, Samarinda, 2020).
11. P. McCullagh and J. A. Nelder. Generalized Linear Models, 2nd Ed. (Chapman and Hall, London, 1989).
12. D. W. Osgood. *J. Quant. Criminol*. 16, 21–43 (2000).
13. J. W. Hardin and J. M. Hilbe. Generalized Linear Models and Extensions. (Stata Press, Texas, 2007).
14. A. C. Cameron and P. K. Trivedi. Regression Analysis of Count Data. (Cambridge University Press, London, 1998).
15. Y. Wang. *Comput. Statist. Data Anal*. 51, 3776–3787 (2007).
16. Y. Pawitan. All Likelihood: Statistical Modelling and Inference Using Likelihood, 1st Ed. (Clarendon Press, Oxford, 2001).