Negative Binomial and Generalized Poisson Regressions in Dengue Hemorrhagic Fever Data at Central Java 2012

M. Al Haris, Muhammad Nur Aidi, Indahwati

Abstract— Poisson regression model, also known as Generalized Linear Model (GLM) was one of the most popular techniques for the analysis of count data. One important assumption for the Poisson regression model was the mean of the distribution must be equal to the variance. Inequality mean and variance led to serious underestimation of standard error and misleading inference for the regression parameters. This problems led an overdispersion. This paper proposed the Negative Binomial and Generalized Poisson regression models as alternative for handling overdispersion. The result shown that based on the test for the dispersion parameter and the goodness-of-fit measure for the total number of Dengue Hemorrhagic Fever morbility Data at Central Java 2012, the Generalized Poisson regression model performed better than the other regression models.

Index Terms— Generalized poisson regression, negative binomial regression, overdispersion, poisson regression.

1 Introduction

Dengue Hemorrhagic Fever (DHF) remains to become the one of the serious public health problems in Indonesia. DHF commonly found in the tropical and subtropical zones that became potensial breeding area for *aedes* mosquitoes, principally *aedes aegypti* [5]. The World Health Organization (WHO) has reported 150,000 cases in Indonesia that led the highest cases in South-East Asia Region, where more 3.5% of the country's population lives in urban areas [10]. DHF was also a serious problem in Central Java Province, it was indicated where all of regencies/cities have been infected by dengue. Incidence Rate (IR) of DHF in Central Java Province in 2012 was 19.29% over 100,000 populations. It increased when compared to 2011 was 15.27% over 100,000 populations and it was still in the national target of <20 over 100,000 populations. Similar with Incidence Rate (IR), the Case Fatality Rate (CFR) of DHF in Central Java Province in 2012 was 1.52% and it was higher than in 2011 (0.93%). This value has already passsed the national target (<1%) [3].

The high number of DHF morbility cases were caused of unstable climate and the amount of rainfall in the rainy season that became *aedes* Mosquitoes potential breeding. It was also supported with mosquito eradication that was not maximal in Central Java society and in recent area DHF caused outbreak [3]. The number of DHF morbility was related some factors based on ephidemiologic triangle. There were three factors playing a role in the infectious diseases and how they spread, namely: the host, the agent and the environment [7].

- *M Al Haris is currently pursuing masters degree program in applied statistics in Bogor Agricultural University, Indonesia, PH +6281392412123. E-mail: alharis3@mail.com*
- *Muhammad Nur Aidi is Lecturer, Departement of Statistics, Bogor Agricultural University, Bogor, Indonesia*
- *Indahwati is Lecturer, Departement of Statistics, Bogor Agricultural University, Bogor,Indonesia*

The agent factor was *Aedes* mosquitoes, principally *Ae. aegypti* transmissing the dengue virus to humans through the bites. The host factor influencing the number of DHF morbility were age, gender, education, employment, imunity, nutritional status, race and bahavior. The environment factor was covering the physical, biological and social environments [7].

The number of DHF morbility was an even that the probability of occurance was small. Modelling that suitable related the number of DHF morbility data with factors that influenced was Poisson regression model [8]. this model requared equality of mean and variance of the dependent variable for each observation. In practice, this assumption was often false since the variance was larger than the mean that was called overdispersion [6]. Overdispersion in Poisson regression model may underestimate the standard error and overstate significance of regression parameters. This condition lead misleading inference about the regression parameters [6]. This paper proposed the Negative Binomial and Generalized Poisson regression models to handle overdispersion.

The purpose of this study was to handle overdispersion in Poisson regression model with Negative Binomial regression and Generalized Poisson regression models and to find the best model from some models produced on the number of DHF morbility data.

2 Research Method

2.1 Data

Data used in this research were secondary data collected from healt departement and Central Bureau of Statistics (BPS) Central Java publication. Unit observation was each regency/city with the total number of DHF morbility each regency/city as dependent variable. The explanatory variables were selected that based on three aspects of the ephidemiologic triangle. There were 10 explanatory variables used which was described in table 1.

TABLE 1
LIST OF VARIABLES

No	Variables	Explanation
1	Y	The total number of DHF morbility in each regency/city
2	X_1	The total number of population in each regency/city
3	X_2	The total number of population under 15 years old in each regency/city
4	X_3	The total number of population density in each regency/city
5	X_4	The total number of population 15 years old that working in each regency/city
6	X_5	The Poverty rate in each regency/city
7	X_6	Mean years of schooling in each regency/city
8	X_7	Clean and healty behavior (PHBS) rate in each regency/city
9	X_8	The total number of health centre service in each regency/city
10	X_9	Percent of children 0-48 months old receiving vaccinations in each regency/city
11	X_{10}	The total number of male population rate in each regency/city

2.2 Methods of Data Analysis

The stages of data analysis in this research involved descriptive analysis, multicollinierity test, overdispersion test, modelling and model evaluation. The detailed stages as follow:

1. Descriptive Analysis
   Descriptive analysis was performed to explore the general description of data patten that aimed to get the appropriate next analysis.
2. Multicollinierity Test
   Multicollinierity on predictor variables should be solved as an assumption for parameter estimation in regression modelling. VIF (Variance Inflation Factor) can be used to detect multicollinierity on predictor variables. Multicollinierity occur if the VIF value was greater than 10. VIF was given by :
   $$\text{VIF} = \frac{1}{(1-R_k^2)} \quad (1)$$
   Where R_k^2 was the coefficient of multiple determinations of the regression obtained by regressing among x_i with the other predictor variables [9].
3. Overdispersion Test
   Overdispersion can be detected by considering deviance or pearson chi-squared value that was devided its degree of freedom. Deviance value was given by :
   $$D = 2\sum_{i=1}^{n}\left\{y_i ln\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i - \hat{\mu}_i)\right\} \quad (2)$$
   While pearson chi-squared value was given by :
   $$Pearson\ \chi^2 = \sum_{i=1}^{n}\frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i} \quad (3)$$
   The existence of overdispersion was indicated if value of deviance or pearson chi-squared that devided by its degree of freedom was greater than 1 [4].
   The alternative test for significance of the overdispersion parameter on the Poisson regression model was used Score Test. The hypothesis was given by
   $$H_0: \Phi = 0$$
   $$H_0: \Phi > 0 \quad (4)$$
   The existence of overdispersion parameter Φ in the Negative Binomial regression or Generalized regression models was confirmed when the null hypothesis was rejected. To evaluated (4), the Score test for overdispersion was given by
   $$S_\phi = \frac{\left(\sum_{i=1}^{n}\left((y_i - \hat{\mu}_i)^2 - y_i\right)\right)^2}{2\sum_{i=1}^{n}\hat{\mu}_i^2} \quad (5)$$
   Where $\hat{\mu}_i$ was the predicted value from the Poisson regression model. under the null hypothesis that the data followed the Poisson regression model. the limiting distribution of the Score value was chi-squared with one degree of freedom, χ_1^2 [2].
4. Modelling of the total number of DHF morbility by Negative Binomial regression.
5. Modelling of the total number of DHF morbility by Generalized Poisson regression.
6. Selection of the best model
   Criteria used to select the best model was Akaike Information Criterion (AIC) value. The smallest AIC value was the best model. AIC value given by :
   $$\text{AIC} = -2lnL(y|\hat{\mu}) + 2p \quad (6)$$
   Where $L(y|\hat{\mu})$ was *log-lielihood* value for the model and p was the number of estimated parameters [4].

3 RESULT AND DISCUSSION

3.1 Descriptive Analysis

Central Java was located in the midle of Java island in Indonesia. Stretches along the equator between 5⁰40′ to 5⁰40′ South Latitude and 108⁰30′ to 111⁰30′ East Longitude. Central Java devided into 29 regencies and 6 cities that spread into 573 sub-district [3].

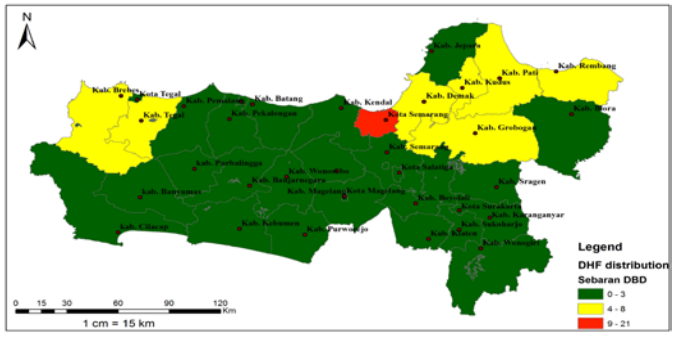
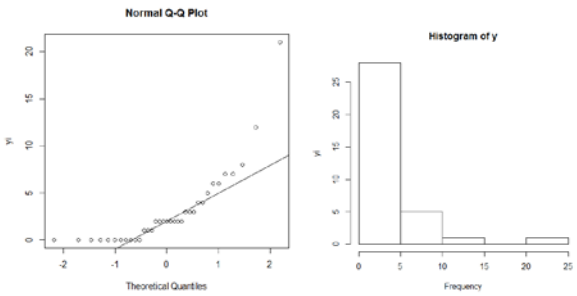


Fig. 1. Map of study area: the number of DHF morbility distribution in Central Java

The distribution of the total number of DHF morbility was showed in Figure 1. This figure indicated that the number of DHF morbility was diverse for each regency/city.



(a) (b)

Fig. 2. (a) The normal quantile-quantile plot and (b) the histogram for the total number of DHF morbility data

Characteristic of the total number of DHF morbility data can be examined by plotting data. Base on Figure 2, the normal Q-Q plot shown that the data distribution permormed skewed negative and did not follow the straight line. The histogram graph also performed the data distribution did not form bell-shaped curve. It indicated that the total number of DHF morbility data did not follow Normal distribution and seemed following Poisson distribution.

3.2 Multicollinierity Test

Modelling that related several explanatory variables must satisfy independencies amog explanatory variables. VIF value of each explanatory variables were shown in Table 2.

TABLE 2
VIF VALUE OF 1O EXPLANATORY VARIABLES

Variables	VIF
X_1	192.261
X_2	113.937
X_3	6.833
X_4	31.100
X_5	3.358
X_6	13.265
X_7	2.176
X_8	5.772
X_9	2.158
X_{10}	3.063

Baded on Table 2, there were several explanatory variables that had VIF value was greather than 10, i.e., x_1, x_2, x_4 and x_6. The hight VIF value indicated there was linear combination among explanatory variables caused multicollinierity [9].

TABLE 3
VIF VALUE OF 7 EXPLANATORY VARIABLES

Variables	VIF
X_1	4.332
X_3	1.912
X_5	1.814
X_7	1.890
X_8	4.021
X_9	1.292
X_{10}	1.818

Process that used to handle multicollinierity on data set was by releasing variables that had VIF value greather than 10 [9]. Several explanatory variables that have been handled was shown in Table 3.

3.3 Overdispersion Test

Poisson regression model requared equality mean and variance (equidispersion) assumption. In practise, several count data displayed that variance exceeded the mean called overdispersion [6]. The investigation of overdispersion in Poisson regression model was shown in table 4.

TABLE 4
THE DEVIANCE AND PEARSON CHI-SQUARED VALUE OF POISSON REGRESSION MODEL

Criterion	Value	df	Value/df
Deviance	78.496	27	2.907
Pearson Chi-Square	99.199	27	3.674

Tabel 4 shown the value of deviance and Pearson chi-squared that devided by its degree of freedom was greather than 1. The result indicated that Poisson regression model was not appropriate used for the total number of DHF morbility data since overdispersion case detected [1].

The computing Score test (5) based on the Poisson regression model was resulted S_ϕ = 4.343 and based on chi-squared table with one degree of freedom χ^2_1 was 3.841. becouse of S_ϕ = 4.343 > χ^2_1 = 3.841, it can be inferred that the Poisson regression model on the total number of DHF morbility data alson significantly occured overdispersion.

The alternative model used to handle overdispersion case on Poisson regression model was Negative Binomial regression and Generalized Poisson regression models [6].

3.4 Negative Binomial Regression Model

TABLE 5
COMPARISON BETWEEN POISSON AND NEGATIVE BINOMIAL REGRESSION MODELS

	Poisson Regression			Negative Binomial Regression		
Variables	Estimate	Std. Error	P-Value	Estimate	Std. Error	P-Value
Intercept	3.22×10^{1}	1.20×10^{1}	0.007*	4.90×10^{1}	2.06×10^{1}	0.017*
X_1	1.90×10^{-6}	5.27×10^{-7}	0.000*	2.11×10^{-6}	8.10×10^{-7}	0.009*
X_3	-7.58×10^{-5}	6.74×10^{-5}	0.261	-1.53×10^{-4}	1.12×10^{-4}	0.170
X_5	-4.05×10^{-2}	2.67×10^{-2}	0.128	-1.47×10^{-3}	4.44×10^{-2}	0.974
X_7	1.02×10^{-2}	1.14×10^{-2}	0.373	6.82×10^{-3}	1.76×10^{-2}	0.699
X_8	-7.89×10^{-3}	2.53×10^{-2}	0.755	-2.48×10^{-2}	3.87×10^{-2}	0.521
X_9	-4.22×10^{-2}	1.64×10^{-2}	0.009*	-5.88×10^{-2}	2.60×10^{-2}	0.024*
X_{10}	-6.07×10^{-1}	2.38×10^{-1}	0.010*	-9.20×10^{-1}	4.03×10^{-1}	0.022*
Dispersi (ϕ)				2.128		
Deviance	78.496			37.402		
Pearson χ^2	99.199			40.313		
Df	27			27		

Negative Binomial regression model was an alternative model to handle count data with overdispersion [1]. The total number of DHF morbility data was estimated using both Poisson regression and Negative Binomial regression models. Table 5 shown the parameter estimates and their standar errors using both Poisson regression and Negative Binomial regression models as comparison.

The total number of DHF morbility data suggested

overdispersion case. The estimation parameter dispersion = 2.128 that indicated overdispersion. Modelling from both Poisson regression and Negative Binomial regression relating 7 explanatory variables resulted x_1, x_9, x_{10} that was significant at 5% level. Estimation parameters from both Poisson regression and Negative Binomial regression were very similar. However, the result in Table 5 clearly indicated that the standart errors from Poisson regression model were overestimate. Consequently, the P-value for testing the significance of each explanatory parameter was generally downward biased for the Poisson regression model. Modelling of the total number of DHF morbility data that obtained based on the Negative Binomial regression was given by

$$\mu_i = \exp(49 + 0.00000211x_1 - 0.000153x_3 - 0.00147x_5 + 0.00628x_7 - 0.0248x_8 - 0.0588x_9 - 0.92x_{10})$$

3.5 Generalized Poisson Regression Model

The other alternative model to handle overdispersion case on the total number of DHF morbility data was Generalized Poisson regression model. The Generalized Poisson regression model was a generalization of standard Poisson regression that was usefull to acomodate overdispersion case on data set [6]. the parameter estimates both Poisson regression and Generalized Poisson regression models as comparison was shown in Table 6.

TABLE 6
COMPARISON BETWEEN POISSON REGRESSION AND GENERALIZED POISSON REGRESSION MODELS

	Poisson Regression			Generalized Poisson Regression		
Variables	Estimate	Std. Error	P-Value	Estimate	Std. Error	P-Value
Intercept	3.22×10^{1}	1.20×10^{1}	0.007*	3.21×10^{1}	1.83×10^{1}	0.079
X_1	1.90×10^{-6}	5.27×10^{-7}	0.000*	2.27×10^{-6}	7.93×10^{-7}	0.004*
X_3	-7.58×10^{-5}	6.74×10^{-5}	0.261	-2.99×10^{-5}	9.70×10^{-4}	0.758
X_5	-4.05×10^{-2}	2.67×10^{-2}	0.128	-2.94×10^{-2}	3.89×10^{-2}	0.449
X_7	1.02×10^{-2}	1.14×10^{-2}	0.373	9.99×10^{-3}	1.72×10^{-2}	0.560
X_8	-7.89×10^{-3}	2.53×10^{-2}	0.755	-1.26×10^{-2}	3.72×10^{-2}	0.735
X_9	-4.22×10^{-2}	1.64×10^{-2}	0.009*	-3.14×10^{-2}	2.47×10^{-2}	0.204
X_{10}	-6.07×10^{-1}	2.38×10^{-1}	0.010*	-6.39×10^{-1}	3.61×10^{-1}	0.077
Dispersi (ϕ)				0.773		
Deviance	78.496			79.961		
Pearson	99.199			86.172		
Df	27			61		

Based on Table 6, The estimated dispersion parameter from the Generalized Poisson regression model was 0.773. the positive value indicated overdispersion that suggested the Poisson regression model was not appropriate to modelling for The total number of DHF morbility data.

The parameter estimates both the Poisson regression and Generalized regression models were quite similar since estimetes from both models were consistent. However, the standart errors from Generalized Poisson regression model were slightly larger than the Poisson regression model. it gave equal inference about the Generalized Poisson regression wiht the Negative Binomial regression modelling.

Difference from Negative Binomial regression model, Generalized Poisson regression model relating 7 explanatory variables resulted only x_1 that was significant at 5% level. Modelling of the total number of DHF morbility data that obtained based on the Generalized Poisson regression was given by

$$\mu_i = \exp(32.1 + 0.00000227x_1 - 0.0000299x_3 - 0.0294x_5 + 0.00999x_7 - 0.0126x_8 - 0.0314x_9 - 0.639x_{10}).$$

3.6 Selection of the Best Model

The criteria for selection of the best model used AIC value. The best model was the model which had the smallest AIC value [4]. AIC value from the all models produced was shown in Table 7.

Base on AIC values in Table 7, the model that had the smallest AIC value was the Generalized Poisson regression model. it was mean that the Generalized Poisson regression model was more appropriate to analyze The total number of DHF morbility data whit several explanatory variables. The model that was resulted given by

TABLE 7
AIC VALUE OF ALL MODELS

Model	AIC
Poisson Regression Model	168.49
Negative Binomial Regression Model	154.34
Generalized Poisson Regression Model	150.70

$$\mu_i = \exp(32.1 + 0.00000227x_1 - 0.0000299x_3 - 0.0294x_5 + 0.00999x_7 - 0.0126x_8 - 0.0314x_9 - 0.639x_{10}).$$

Modelling the total number of DHF morbility data by Generalized Poisson regression model was only related explanatory the total number of population in each regency/city (x_1). The obtained model can be explained that when every one inhabitant was added in each regency/city, it will increase the expectation of the total number of DHF morbility exp (0.00000227) = 1. 00000227 times with the other variables assumed to be constant.

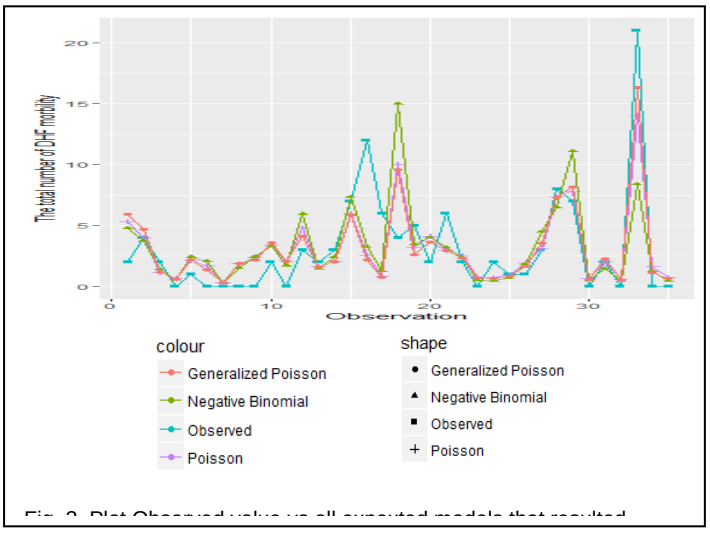


Fig. 3. Plot Observed value vs all expected models that resulted

Figure 3. was shown that the expected value by the Generalized Poisson regression model was more closed with observed value than the other models. It supported that the Generalized Poisson regression model was more appropriate

to analyze the total number of DHF morbility data.

4 Conclusion

With the growing population in Central Java, the total number of DHF morbility continued to rise. The classical model that related the total number of DHF morbility data with several explanatory variables was Poisson regression model. However, the modelling with the Poisson regression occured overdispersion case. This research suggested the Negative Binomial regression and the Generalized Poisson regression models as an alternative to overcome overdispersion case. The result shown that the Poisson regression, Negative Binomial regression and Generalized Poisson regression models produced similar estimated for parameter estimates, but the standard errors for Poisson regression model was smaller than the other models. Therefore, the Poisson regression model overestimated the significance of the regression parameters caused the presence of overdispersion. Base on the AIC value for all models produced, the Generalized Poisson regression model was the best model and more appropriate to analyze The total number of DHF morbility data with several explanatory variables.

Reference

[1] A. Melliana, "The Comparison of Generalized Poisson Rgression and Negative Binomial Regression Methods in Overcoming Overdispersion", *International Journal of Scientific & Technology Research*, 2013, Vol 2, Issue 8.

[2] D.T. Molla and B. Muniswamy, "Power of Test for Overdispersion Parameter in Negative Binomial Regression Model", *IOSR Journal of Mathematics*, 2012, pp. 29-36.

[3] Health Department of Central Java, *Profil Kesehatan Provinsi Jawa Tengah 2012*, 2012, Semarang: Health Department of Central Java.

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

[5] Ministry of Health, *Profil Kesehatan Indonesia Indonesia 2012*, 2012, Jakarta: Ministry of Health.

[6] N. Ismail and A.A. Jemain, "Handling Overdispersion with Negative Binomial and Generalized Poisson Regression Models", 2007, Virginia: *Casualty Actuarial Society Forum*.

[7] N.P. Widodo, *Faktor-Faktor yang Berhubungan dengan Kejadian Demam Berdarah Dengue (DBD) di Kota Mataram Provinsi Nusa Tenggara Barat Tahun 2012*", 2012, Depok: Universitas Indonesia, (THESIS).

[8] P. McCullagh and J.A. Nelder, *Generalized Linear Model (2nd Edition)*, 1989 London: Chapman and Hall.

[9] R.H. Myers, *Classical and Modern Regression with Applications (2nd Edition)*, 1990, New York: PSW-KENT.

[10] WHO, "Dengue: Guidlines for Diagnosis, Treatment, Prevention and Control New Edition", 2009, Genewa: World Health Organization.