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 potential 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 epidemiologic 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].

The agent factor was Aedes mosquitoes, principally Ae. aegypti transmitting the dengue virus to humans through the bites. The host factor influencing the number of DHF morbility were age, gender, education, employment, immunity, nutritional status, race and behavior. 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 required 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 health 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 epidemiologic triangle. There were 10 explanatory variables used which was described in table 1.
2.2 Methods of Data Analysis

The stages of data analysis in this research involved descriptive analysis, multicollinearity 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 pattern that aimed to get the appropriate next analysis.

2. **Multicollinearity Test**
   Multicollinearity on predictor variables should be solved as an assumption for parameter estimation in regression modelling. VIF (Variance Inflation Factor) can be used to detect multicollinearity on predictor variables. Multicollinearity occur if the VIF value was greater than 10. VIF was given by:
   \[
   VIF = \frac{1}{(1-R_i^2)}
   \]
   Where \( R_i^2 \) 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 devied its degree of freedom. Deviance value was given by:
   \[
   D = 2 \sum_{i=1}^{n} \left( y_i \ln \left( \frac{\hat{\mu}_i}{\hat{\mu}_i} \right) - (y_i - \hat{\mu}_i) \right)
   \]
   While pearson chi-squared value was given by:
   \[
   Pearson \chi^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}
   \]
   The existence of overdispersion was indicated if value of deviance or pearson chi-squared that devied by its degree of freedom was greater than 1 [4].

   The alternative test for significance of the overdispersion parameter \( \Phi \) 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:
   \[
   s_p = \frac{\left( \sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2 / \hat{\mu}_i \right)^2}{2 \sum_{i=1}^{n} R_i^2}
   \]
   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, \( \chi^2 \) [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:
   \[
   AIC = -2\ln L(y|\hat{\beta}) + 2p
   \]
   Where \( L(y|\hat{\beta}) \) was log-likelihood 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 5°40' to 5°40' South Latitude and 108°30' to 111°30' East Longitude. Central Java devided into 29 regencies and 6 cities that spread into 573 sub-district [3].

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.
Characteristic of the total number of DHF morbility data can be examined by plotting data. Based on Figure 2, the normal Q-Q plot shown that the data distribution performed 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 among explanatory variables. VIF value of each explanatory variables were shown in Table 2.

**TABLE 2**

<table>
<thead>
<tr>
<th>Variables</th>
<th>VIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_1$</td>
<td>192.261</td>
</tr>
<tr>
<td>$X_2$</td>
<td>113.937</td>
</tr>
<tr>
<td>$X_3$</td>
<td>6.833</td>
</tr>
<tr>
<td>$X_4$</td>
<td>31.100</td>
</tr>
<tr>
<td>$X_5$</td>
<td>3.358</td>
</tr>
<tr>
<td>$X_6$</td>
<td>13.265</td>
</tr>
<tr>
<td>$X_7$</td>
<td>2.176</td>
</tr>
<tr>
<td>$X_8$</td>
<td>5.772</td>
</tr>
<tr>
<td>$X_9$</td>
<td>2.158</td>
</tr>
<tr>
<td>$X_{10}$</td>
<td>3.063</td>
</tr>
</tbody>
</table>

Based on Table 2, there were several explanatory variables that had VIF value greater than 10, i.e., $X_1$, $X_2$, $X_4$ and $X_6$. The high VIF value indicated there was linear combination among explanatory variables caused multicollinierity [9].

**TABLE 3**

<table>
<thead>
<tr>
<th>Variables</th>
<th>VIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_1$</td>
<td>4.332</td>
</tr>
<tr>
<td>$X_3$</td>
<td>1.912</td>
</tr>
<tr>
<td>$X_5$</td>
<td>1.814</td>
</tr>
<tr>
<td>$X_7$</td>
<td>1.890</td>
</tr>
<tr>
<td>$X_9$</td>
<td>4.021</td>
</tr>
<tr>
<td>$X_{10}$</td>
<td>1.292</td>
</tr>
</tbody>
</table>

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

### 3.3 Overdispersion Test

Poisson regression model required 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**

<table>
<thead>
<tr>
<th>Deviance and Pearson Chi-Squared Value of Poisson Regression Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Criterion</td>
</tr>
<tr>
<td>Deviance</td>
</tr>
<tr>
<td>Pearson Chi-Square</td>
</tr>
</tbody>
</table>

Tabel 4 shown the value of deviance and Pearson chi-squared that devied by its degree of freedom was greater 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 $\chi^2 = 4.343$ and based on chi-squared table with one degree of freedom $\chi^2 = 3.841$. Because of $\chi^2 = 4.343 > 3.841$, 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 Binominal regression and Generalized Poisson regression models [6].

### 3.4 Negative Binomial Regression Model

**TABLE 5**

<table>
<thead>
<tr>
<th>Comparison Between Poisson and Negative Binominal Regression Models</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variables</td>
</tr>
<tr>
<td>Intercept</td>
</tr>
<tr>
<td>$X_1$</td>
</tr>
<tr>
<td>$X_2$</td>
</tr>
<tr>
<td>$X_3$</td>
</tr>
<tr>
<td>$X_4$</td>
</tr>
<tr>
<td>$X_5$</td>
</tr>
<tr>
<td>$X_6$</td>
</tr>
<tr>
<td>$X_7$</td>
</tr>
</tbody>
</table>

Negative Binominal 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 Binominal regression models. Table 5 shown the parameter estimates and their standard errors using both Poisson regression and Negative Binominal regression models as comparison.

The total number of DHF morbility data suggested
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 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 with several explanatory variables. The model that was resulted given by

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

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.

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 estimates from both models were consistent. However, the standard errors from Generalized Poisson regression model were slightly larger than the Poisson regression model. It gave equal inference about the Generalized Poisson regression with the Negative Binomial regression modelling.

Difference from Negative Binomial regression, 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
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 occurred 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