Modeling The Number of Flood Occurrence in Indonesia in 2015 Using Poisson Regression Model

Okky Savitri Febiyani, Laksmi Prita Wardhani
Faculty of Mathematics, Computing, and Data Sciences, Sepuluh Nopember Institute of Technology
Arief Rahman Hakim street, Surabaya 60111 Indonesia
E-mail: okky.savitri.febiyani@gmail.com

Abstract. The Poisson regression model is one of the regression models that can show the relationship between the Poisson distributed response variable with predictor variables through the regression parameter. In this paper, the number of flood occurrences in Indonesia in 2015 which is the count data using a Poisson regression model. To obtain the Poisson regression model of the number of flood occurrence are conducted using multicollinearity test on predictor variables, estimation of regression parameters, parameter test, and goodness of fit model test. The method used to estimate parameter values is the Maximum Likelihood Estimation method that is obtained from the iteration results using the Newton-Raphson iterative method. Based on the results of data processing, the Poisson regression model that is obtained for the number of flood occurrences shows that the average of the number of flood occurrences in Indonesia is affected by the number of tornado occurrences, the number of rainy days, and the sun exposure.

1. Introduction

Indonesia is a country that has a tropical climate caused by Indonesia’s geographical position on the equator. Thus Indonesia has a constant temperature in almost all regions of Indonesia with an average temperature of 80°F throughout the year. Humidity in this country is generally very high with varying rainfall levels due to monsoon. In addition, the topography of Indonesia’s islands is mountainous and interspersed with low land around the coast. With these characteristics, Indonesia is vulnerable to various natural disasters such as floods, droughts, tornado, landslides, tidal waves, etc. [1].

According to the National Disaster Management Agency (BNPB), flooding is an event or situation in which an area or land is submerged due to increased water volume. Based on data recorded by BNPB, between 2000 and 2015, floods contributed the highest disaster in Indonesia, which was around 32%, with a total incidence of 6.416. In addition, the disaster has also caused casualties, property losses and damage to buildings.

Flooding can be caused by several natural factors, such as extreme rainfall and the presence of rising tides of seawater [2]. Therefore, it is necessary to conduct research on
Applied Mathematics

floods to obtain any factors that significantly influence the number of flood occurrences simultaneously in an effort to reduce the number of casualties and property losses incurred.

Poisson regression is one regression model used to model the response variable in the form of count data. Poisson regression model describes the relationship between response variables and predictor variables through regression parameters [3]. In this research, a Poisson regression model is applied to model the number of flood events in Indonesia in 2015 which is count data with factors that are thought to influence the number of flood occurrences. The factors used are the number of drought occurrences, the number of tornado occurrences, the number of tidal events, the amount of rainfall, the number of rainy days, wind speed, humidity, average temperature, air pressure, and sun exposure.

2. Literature Review

2.1. Poisson Distribution

The experiment which results the numeric number from random variable $Y$ that happens during the determined interval of time or area is called Poisson experiment. The number of $Y$ resulted during the Poisson experiment is called Poisson random variable, and its probability distribution is called Poison distribution [4].

Poisson distribution is appropriate for dependent variable with non-negative integer value which is $0, 1, 2, 3, \ldots$, along with $\lambda$ parameter, $\lambda > 0$. The probability function for the Poisson distribution is as follows [5]:

$$f(y|\lambda) = \frac{e^{-\lambda}\lambda^y}{y!}, y = 0, 1, 2, 3, \ldots$$  \hspace{1cm} (1)

where:

$y$ : the random data of the number of flood incident

$\lambda$ : the average number of flood incident

The simplest distribution for data count is the Poisson. In this research, the number of flood occurrences is a count data so that it can be assumed that the number of flood occurrences has a Poisson distribution. Therefore equation (1) which is a probability density function of the number of flood occurrences will be used to obtain the estimated value of the regression parameter which states the relationship between the number of flood occurrences as the response variables with predictor variables.

2.2. Poisson Regression

The number of flood incidents in Indonesia happened in 2015 as count data was used in this research. The simplest distribution for count data is through Poisson distribution with predictor variables which are considered effective. The model of Poisson regression
Applied Mathematics

is written as follows [6] :

\[ \ln \lambda_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ki} \]  

(2)

where :

\( \lambda_i \) : the average number of flood incidents on \( i \)-observation, \( i = 1, 2, \ldots, n \)

\( \beta_p \) : \( p \)-regression parameter, \( p = 0, 1, 2, \ldots, k \)

\( x_{pi} \) : the data of \( p \)-predictor variable on \( i \)-observation

2.3. Multicollinearity

Multicollinearity is a condition in which there is a linear relationship among predictor variables from the regression model. The existence of multicollinearity is verified through calculating the value of the Variance Inflation Factor (VIF) formulated as follows [4]:

\[ VIF = \frac{1}{1 - R^2_j} \]  

(3)

with \( R^2_j \) is as determination coefficient of predictor variable \( X_j \) with other predictor variables. If the determination coefficient is almost 0, the value of VIF will be close to 1, which indicates that there is no multicollinearity. In the rule of VIF, the value which exceeds 10 indicates the number of multicollinearity problems [3].

2.4. The Estimation of Regression Parameter

Maximum Likelihood Estimation (MLE) method refers to a method used to estimate the regression parameter whose distribution is already known. The estimation of the parameter value is obtained by maximizing its likelihood function which is defined as follows [5] :

\[ L(y; \lambda) = \prod_{i=1}^{n} f(y_i; \lambda_i) = f(y_1; \lambda_1)f(y_2; \lambda_2) \ldots f(y_n; \lambda_n) \]  

(4)

where \( f(y_i; \lambda_i) \) is the probability function of its response variable.

As the Poisson distribution is a member of exponential family, in making the calculation easier, the likelihood function is changed into logarithmic form. Thus, the function of log-likelihood as Poisson regression model that has poisson-distributed response variable is

\[ l(\lambda) = \sum_{i=1}^{n} (-\lambda_i + y_i \ln \lambda_i - \ln y_i!) \]  

(5)

As \( \lambda_i = \exp(x'_i\beta) \), the equation (5) can be repeatedly written as follows :

\[ l(\beta) = \sum_{i=1}^{n} (-\exp(x'_i\beta) + y_i(x'_i\beta) - \ln y_i!) \]  

(6)
Applied Mathematics

In getting the value of $\hat{\beta}$ which maximizes the function of equation (6), Newton-Raphson iterative method was used by fulfilling the condition on the first derivative as follows:

$$\frac{\partial l(\beta)}{\partial \beta} = 0 \quad (7)$$

Newton-Raphson Iteration Formula is written as follows:

$$\hat{\beta}_{s+1} = \hat{\beta}_s - \hat{H}_s^{-1}\hat{g}_s \quad (8)$$

where:
- $\hat{\beta}_s$ : the vector of regression parameters on $s$-iteration, $s = 0, 1, 2, \ldots$
- $\hat{\beta}_{s+1}$ : the vector of regression parameters on $s + 1$ iteration
- $\hat{g}_s$ : evaluated gradient vector on $\hat{\beta}_s$
- $\hat{H}_s$ : evaluated Hessian Matrix on $\hat{\beta}_s$

2.5. The Testing of Poisson Regression Parameter

In investigating the effect given to each predictor variable, parameter test was carried out simultaneously and partially. The testing of Poisson regression parameter simultaneously used Likelihood Ratio (LR) test. The hypothesis used to test the parameter simultaneously was as follows:

Hypothesis :
- $H_0 : \beta_1 = \beta_2 = \beta_3 = \cdots = 0$ (there is no variable which has significant effect)
- $H_1 : \beta_j \neq 0, j = 1, 2, 3, \ldots, k$ (there is at least one variable which has significant effect)

The test statistics used was:

$$\Lambda = 2 \left[ -\sum_{i=1}^{n} \exp(x'_i\hat{\beta}) + \sum_{i=1}^{n} y_i(x'_i\hat{\beta}) + ne^{\hat{\beta}_0} - \hat{\beta}_0 \sum_{i=1}^{n} y_i! \right] \quad (9)$$

The criteria of the test is $H_0$ is rejected if $\Lambda > \chi^2(\alpha, p)$ in which $p$ is the number of parameters [8].

Meanwhile, the partial test of Poisson regression parameter used hypothesis as the following:

Hypothesis :
- $H_0 : \beta_j = 0, j = 1, 2, \ldots$ (the effect of $j$-variable is not significant)
- $H_1 : \beta_j \neq 0$ (the effect of $j$-variable is significant)

The test statistics used is:

$$W_j = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \quad (10)$$

The criteria of the test is $H_0$ is rejected if $|W_j| > t(\alpha, n-1)$, in which $\alpha$ is the significant level and $n$ is the number of observations. It means that there is an effect of tested predictor variables on response variables [4].
Applied Mathematics

2.6. The Goodness of Fit Testing of Poisson Regression Model

The goodness of fit of the Poisson Regression Model can be determined by using the deviance statistics. The following is the hypothesis formulation for the suitability testing of the Poisson regression model.

Hypothesis:

\[ H_0 : y_i = \lambda(x_i; \beta), \quad i = 1, 2, 3, \ldots, n \]
\[ H_1 : y_i \neq \lambda(x_i; \beta) \]

The test statistics used is Deviance. The Deviance formula for Possion regression model is formulated as follows:

\[
\text{Deviance} = 2 \sum_{i=1}^{n} \left[ y_i \ln \left( \frac{y_i}{\hat{\lambda}_i} \right) - (y_i - \hat{\lambda}_i) \right] v
\]  

(11)

The criteria of the test is \( H_0 \) is accepted at \( \alpha \) significance level if \( \text{Deviance} > \chi^2_{(n-p,\alpha)} \), where \( n \) is the number of observations and \( p \) is the number estimated parameters [8].

3. Research Methodology

The secondary data from National Board for Disaster Management and Central Bureau of Statistics regarding the number of flood occurrences in Indonesia happened in 2015 was taken as the predictor variables, which were the number of drought, tornado, tidal wave, rainfall, rainy day, wind speed, humidity, average temperature, air pressure and sun exposure. The notation of each variable can be seen in Table 1.

| Notation | Explanation                                      |
|----------|---------------------------------------------------|
| Y        | The number of flood occurrences                   |
| X_1      | The number of drought                             |
| X_2      | The number of tornado                             |
| X_3      | The number of tidal waves                         |
| X_4      | Rainfall                                          |
| X_5      | Rainy day                                         |
| X_6      | Wind speed                                        |
| X_7      | Humidity                                          |
| X_8      | Average temperature                               |
| X_9      | Air pressure                                      |
| X_{10}   | Sun exposure                                      |

4. Analysis and Discussion

This section provides the formulation of the Poisson regression model for the number of flood incidents covering the test of multicollinearity, parameter, and goodness of fit model.
4.1. The Formation of Poisson Regression Model

The Poisson regression model of flood incidents used 10 predictor variables on 34 observations by following the equation (2) which was written as follows:

$$\ln \lambda_i = \begin{pmatrix} 1 & x_{i1} & x_{i2} & \cdots & x_{i10} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{10} \end{pmatrix} = (x'_i \beta)$$ \hspace{1cm} (12)

for $i = 1, 2, \ldots, 34$.

The value of $\beta$ on equation (2) was gained through Maximum Likelihood Estimation (MLE) method. In accordance with the equation (7), it was obtained:

$$\frac{\partial l(\beta)}{\partial \beta} = \sum_{i=1}^{n} [(y_i - \exp(x'_i \beta))x_{ji}] = 0$$ \hspace{1cm} (13)

The equation (13) was an equation which was not in the closed form as it did not show the analytical solution, so it was necessary to apply the iterative method of Newton-Raphson through the iteration formula that correspond to the equation (8). The gradient vector and Hessian matrix were shown in the following:

$$\hat{g}_s = \frac{\partial l(\beta)}{\partial \beta} \bigg|_{\hat{\beta}_s} \text{ and } \hat{H}_s = \frac{\partial^2 l(\beta)}{\partial \beta \partial \beta'} \bigg|_{\hat{\beta}_s}$$ \hspace{1cm} (14)

According to the equation (14), $\hat{g}_s$ and $\hat{H}_s$ could be described as follows:

$$\hat{g}_s = \begin{pmatrix} \frac{\partial l(\beta)}{\partial \beta_0} \\ \frac{\partial l(\beta)}{\partial \beta_1} \\ \vdots \\ \frac{\partial l(\beta)}{\partial \beta_{10}} \end{pmatrix}, \hat{H}_s = \begin{pmatrix} \frac{\partial^2 l(\beta)}{\partial \beta_0^2} & \frac{\partial^2 l(\beta)}{\partial \beta_0 \partial \beta_1} & \cdots & \frac{\partial^2 l(\beta)}{\partial \beta_0 \partial \beta_{10}} \\ \frac{\partial^2 l(\beta)}{\partial \beta_1 \partial \beta_0} & \frac{\partial^2 l(\beta)}{\partial \beta_1^2} & \cdots & \frac{\partial^2 l(\beta)}{\partial \beta_1 \partial \beta_{10}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 l(\beta)}{\partial \beta_{10} \partial \beta_0} & \frac{\partial^2 l(\beta)}{\partial \beta_{10} \partial \beta_1} & \cdots & \frac{\partial^2 l(\beta)}{\partial \beta_{10}^2} \end{pmatrix}$$

To obtain $\hat{g}_s$, the first derivative of log-likelihood was necessary to find on $\beta_j$, $j = 0, 1, 2, \ldots, 10$. The formulation obtained was:

$$\frac{\partial l(\beta)}{\partial \beta_j} = -\sum_{i=1}^{n} \exp(x'_i \beta)x_{ji} + \sum_{i=1}^{n} y_i x_{ji}, \hspace{1cm} j = 0, 1, 2, \ldots, 10$$ \hspace{1cm} (15)

To obtain $\hat{H}_s$, the second derivative should be found from the function of log-likelihood on $\beta_j$ and $\beta_c$, $c, j = 0, 1, 2, \ldots, 10$. The formulation was as follows:

$$\frac{\partial^2 l(\beta)}{\partial \beta_j \partial \beta_c} = -\sum_{i=1}^{n} \exp(x'_i \beta)x_{ji}x_{ci}, \hspace{1cm} j, c = 0, 1, 2, \ldots, 10$$ \hspace{1cm} (16)

Then, the equation (15) was substituted to $\hat{g}_s$ and the equation (16) was to $\hat{H}_s$ in doing the iteration through the equation (14).
Applied Mathematics

4.2. Multicollinearity Test

Before the Poisson regression parameter was estimated from the model of the number of flood incidents, the multicollinearity test was carried out on predictor variables which were suspected effective. To determine whether or not there was multicollinearity between predictor variables, the calculation of VIF value was taken as in the equation (3). Based on each VIF value on predictor variable shown in Table 2, each VIF value on each variable \( X_j, j = 1, 2, \ldots, 10 \) was less than 10. It indicates that there was no multicollinearity on each predictor variable \( X_j, j = 1, 2, \ldots, 10 \).

4.3. The Poisson Regression Model of Flood Occurrence

Based on the result of data tabulation with R application, the estimated parameter obtained was attached on Table 3.

\[ \begin{array}{c|c}
\text{Parameter} & \text{Estimation} \\
\hline
\beta_0 & 4.686 \\
\beta_1 & -0.1107 \\
\beta_2 & 0.02154 \\
\beta_3 & 0.08101 \\
\beta_4 & -0.00009358 \\
\beta_5 & -0.005320 \\
\beta_6 & 0.07571 \\
\beta_7 & -0.01644 \\
\beta_8 & -0.2051 \\
\beta_9 & 0.008042 \\
\beta_{10} & -0.04644 \\
\end{array} \]

Therefore, the temporary model of flood occurrences in Indonesia which was in 2015
used Poisson regression model as follows:

$$\lambda_i = \exp(4.686 - 0.1107x_{1i} + 0.02154x_{2i} + 0.08101x_{3i} - 0.00009358x_{4i} - 0.005320x_{5i} +$$
$$0.07571x_{6i} - 0.01644x_{7i} - 0.2051x_{8i} + 0.008042x_{9i} - 0.04644x_{10i})$$

(17)

4.4. The testing of Poisson Regression Parameter

The parameter test was conducted simultaneously and partially. The following are the hypothesis, test statistics and the criteria used for both simultaneous and partial parameter test.

(i) Simultaneous Test

The test of the simultaneous parameter was used to know the relationship of predictor variable simultaneously on the number of flood incidents.

Hypothesis :

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_{10} = 0 \text{ (there is no variable which has significant effect)}$$

$$H_1 : \beta_j \neq 0, \ j = 1, 2, 3, \ldots, 10 \text{ (there is at least one variable which has significant effect)}$$

Test statistics :

In relation to the test statistics on the equation (9), it was obtained that at least there was a predictor variable that has significant effect on the number of flood occurrences.

(ii) Partial Test

The test of partial parameters was used to know the predictor variables which have significant effect.

Hypothesis :

$$H_0 : \beta_j = 0, \ j = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 \text{ (the effect of } j\text{-variable is not significant)}$$

$$H_1 : \beta_j \neq 0 \text{ (the effect of } j\text{-variable is significant)}$$

Test statistics :

In accordance with the equation (10), it was obtained that the effective variables were $X_2$, $X_5$, and $X_{10}$. The variable that had no significant effect was eliminated and it was then applied by the modelling of the number of flood with three predictor variables which were $X_2$, $X_5$, and $X_{10}$. The following was the value of estimated parameters from a new model.

Therefore, a new model of the number of flood occurrences could be summed up as follows:

$$\lambda_i = \exp(5.880176 + 0.019770x_{2i} - 0.006001x_{5i} - 0.043784x_{10i})$$

(18)

The new model of the number of flood incidents in Indonesia in 2015 was shown on the equation (18) as it pointed two negative regression parameters, which were the number
Applied Mathematics

Table 4. The Parameter Estimation of Poisson Regression Model on The Number of Flood Occurrences with $X_2, X_5, X_{10}$ as Predictor Variables

| Parameter | Estimation | Std. Error |
|-----------|------------|------------|
| $\beta_0$ | 5.880176   | 0.469243   |
| $\beta_2$ | 0.019770   | 0.001085   |
| $\beta_5$ | -0.006001  | 0.001611   |
| $\beta_{10}$ | -0.043784 | 0.005404   |

of rainy days ($X_5$) and the sun exposure ($X_{10}$). Meanwhile, the number of tornado ($X_2$) had positive coefficient.

After obtaining the estimation value of $\beta_0$, $\beta_2$, $\beta_5$, and $\beta_{10}$ on the new model, the parameter test was done partially to find out whether or not the predictor variables $X_2$, $X_5$, and $X_{10}$ gave an effect.

Hypothesis:

$H_0 : \beta_j = 0$, $j = 2, 5, 10$ (the effect of $j$-variable is not significant)

$H_1 : \beta_j \neq 0$ (the effect of $j$-variable is significant)

Test statistics:

In accordance with the equation (10), it was obtained that three predictor variables ($X_2, X_5, X_{10}$) had significant effect on the number of flood occurrence.

4.5. The Test of Goodness of Fit Model

Based on the test of the parameters, a new model was obtained along with three variables that significantly affected the number of flood occurrences. Then, the test on the model was conducted to find out whether or not the model is suited for the number of flood occurrences data in Indonesia in 2015. The following hypothesis was used to test the suitability model.

Hypothesis:

$H_0 : y_i = \lambda(x_i; \beta), \ i = 1, 2, \ldots, 34$

$H_1 : y_i \neq \lambda(x_i; \beta)$

Test Statistics:

Based on the test statistics in the equation (11), it was obtained that the new model on the equation (18) was suitable with the actual data $Y$. From this model, the number of flood occurrences in Indonesia occured in 2015 was influenced by $\exp(5.880176)$ with the increased $\exp(0.019770)$ of the number of tornado occurrences ($x_2$), the reduced $\exp(0.006001)$ of the number of rainy days ($x_5$), and the decreased $\exp(0.043784)$ of sun exposure ($x_{10}$) on each $i$-observation.


Applied Mathematics

5. Conclusion

According to the modelling results of the number of flood occurrences in Indonesia occurred in 2015, the regression parameter $\beta_0$ was 5.880176, $\beta_2$ was 0.019770, $\beta_5$ was $-0.006001$ and $\beta_{10}$ was $-0.043784$, which partially explained the significant effect of the number of tornado, rainy day and sun exposure on the number of flood occurrences. In short, the Poisson regression model on the number of flood occurred in 2015 was written as follows:

$$\lambda_i = \exp(5.880176 + 0.019770 x_{2i} - 0.006001 x_{5i} - 0.043784 x_{10i})$$

The model stated that the average number of flood occurrences in Indonesia, specifically in 2015 on each $i$-observation was $\exp(5.880176)$ with the increased $\exp(0.019770)$ of the number of tornado ($x_{2i}$), the reduced $\exp(0.006001)$ of the number of rainy days ($x_{5i}$) and the decreased $\exp(0.043784)$ of sun exposure ($x_{10i}$) on each $i$-observation.

References

[1] Center for Excellence in Disaster Management and Humanitarian Assistance 2015 Indonesia Disaster Management Reference Handbook, Center for Excellence in Disaster Management and Humanitarian Assistance, U.S.
[2] National Agency for Disaster Management 2012 Buku Saku Tanggap Tangkas Tanggah Menghadapi Bencana (Indonesia : National Agency for Disaster Management)
[3] Gujarati, D.N. 2003 Basic Econometrics 4th ed (New York : McGraw-Hill/Irwin)
[4] James, G., et al. 2013 An Introduction to Statistical Learning (New York : Springer Science + Business Media)
[5] Cameron, A.C. and Trivedi, P.K. 2005 Microeconometrics (New York : Methods and Applications, Cambridge University Press)
[6] Agresti, A. 2002 Categorical Data Analysis 2nd ed (New Jersey : John Wilry & Sons, Inc.)
[7] Dobson, A.J. and Barnett, A.G. 2008 An Introduction to Generalized Linear Models 3rd ed (USA : Chapman & Hall/CRC)
[8] Cameron, A.C. and Trivedi, P.K. 1998 Regression Analysis of Count Data (New York : Cambridge University Press)