APPLICATION OF POISSON REGRESSION IN MODELING THE NUMBER OF DENGUE CASES SUFFERERS BASED ON SOCIO-DEMOGRAPHIC AND TEMPERATURE IN INDONESIA

Oscar [2], Pamela [3], F Ranny [4], ES Nugraha [4]
School of Business, Study Program of Actuarial Science, President University, Cikarang, Bekasi 17550, Indonesia
[2]*oscar@student.president.ac.id*
[3]*pamela@student.president.ac.id*
[4]*ranny.febrianti@student.president.ac.id*
[1]*edwin.nugraha@president.ac.id*

ABSTRACT
Indonesia is one of the countries with high dengue cases. Dengue dengue is an acute disease with clinical manifestations of the bleeding that causes shock that leads to death. Dengue or *Demam Berdarah* is a type of disease caused by one out of four dengue viruses. This disease is highly contagious. The source of the infection is originally coming from *Aedes aegypti* and *Aedes albopictus* mosquito bite. Some of the factors that are known to influence dengue cases include home environment, biological environment, and social environment. With Poisson regression, we connect other factors that have possible risk in dengue cases. In these cases, we used data on dengue cases 2015 in Indonesia. Our analysis with help of R software shows that poverty and temperature will make a significant contribution to the number of dengue 2015 cases in Indonesia. Every 1% increase in the percentage of poverty will cause a decrease in dengue cases to exp (-0.04449) = 0.9564 times from the initial DHF case assuming other predictor variables are constant. Every 1^0 Celsius increase in temperature will cause an increase in DHF cases to exp (0.206) = 1,306 times from the initial dengue case assuming the other predictor variables are constant. This research is expected to help the health sector and government in dengue control in Indonesia.

Keywords: *Dengue, Poisson Regression, Factors, Health, Indonesia.*

1. Introduction

Indonesia is a country with a tropical climate that is good for animal life and plants. It makes Indonesia an excellent place to see disease progression, especially disease vector transmission. Fever bleeding (DHF) Dengue is a disease caused by the dengue virus with vector transmission (Sukohar, 2014). Dengue dengue is an acute disease with clinical manifestations of the bleeding that causes shock that leads to death. Dengue or *Demam Berdarah* is a type of disease caused by one out of four dengue viruses. This disease is highly contagious. The source of the infection is originally coming from the *Aedes aegypti* and *Aedes albopictus* mosquito bite. Both types of mosquito are biting in the morning until approaching sunset. The transmission itself occurs when a mosquito bites and sucks humans' blood that has been infected by the dengue virus. When the mosquito is biting another person, the virus will be spread. Dengue is being triggered by several factors such as having had a previous dengue virus infection, living and travelling to the tropics region, and Infants, children, the elderly, and people with weakened immunity (Widiyanto, 2007).

Possible risk factors that may affect the incidence of dengue include: home environment (distance house, house layout, container type, height place), climate, biological environment, and social environment. Distance between houses affects the spread of mosquitoes from one house to house. The closer the distance between houses, the easier the mosquitoes spread to the house next door. Homebuilding materials, construction of the house, the colour of the walls and the arrangement of the items in the house cause the house to be liked or disliked by a mosquito (Prasetyani, 2015). Various infectious diseases prove that the housing conditions were jostling and slums have the possibility of the disease being attacked is greater.

In Indonesia, as of December 3, 2020, dengue cases were spread into 472 districts/cities in 34 Provinces. The deaths caused by this disease occurred in 219 districts/cities. On November 30, 2020, there were 51 additional cases and 1 death due to dengue. The Ministry of Health Didi Budijanto appealed to the public

to implement *Pemberantasan Sarang Nyamuk (PSN)3M Plus* (Desniawati, 2014). The first M refers to *Menguras*, which is an activity to clean/drain places that are often used as water reservoirs such as bathtubs, jugs, water toren, drums, and others. The next M is *Menutup,* which is an activity to close tightly the water reservoirs such as bathtubs and drums. Last M has to be *Memanfaatkan Kembali* or reusing waste of used goods that are economically valuable (recycled).

In previous studies, it was stated that the negative binomial is the best model that can model the relationship between the influence of climate factors on the number of DBD sufferers in the city of Bogor (Sihombing & Sundari 2014). However, in this study, the coverage area is all over Indonesia, and the best model according to our data is also different from previous studies. (Prasetyowati, H., et.all, 2015) has been conducted research in 27 districts of West Java Province with the same method as this study, focusing on one of the factors driving the increase in dengue cases in public environmental health conditions settlements, revealed that the components of environmental health that affect the incidence of dengue are the way of handling waste, disposing of waste water, and draining the bath. The previous study was comparing which model is the best between Neural Networks and Poisson Regression in cases of dengue (DBD) in Surabaya. The result proves that the best model uses a neural network model, but we want to prove that using Poisson regression also gives better results in this study. In terms of coverage, the previous researchers covered the city of Surabaya, while we covered the whole of Indonesia. (Pradhani, 2016)

Dr. Tedjo Sasmono, Head of Dengue Research Unit at the Eijkman Institute of Molecular Biology, stated that "The extermination of dengue is highly difficult since it comes up with many factors. Starting from the climate, mosquito vector and its population, and immunity". Our group has different points of view on the factors affecting dengue. Those factors are life, sociodemographic, and temperature. As it is followed by our research purpose or research questions:

1. Is there a relationship between population, poor population, temperature and dengue cases in Indonesia?
2. To obtain the best model using Poisson Regression.

2. Literature Review

2.1. Poisson Distribution

Poisson distribution is a likelihood conveyance that expresses the likelihood of various occasions happening in a period (Junadi, 1995). The Poisson dissemination can be utilized to communicate occasions in explicit units or periods over the long haul. Utilization of the Poisson Distribution as the reason for Poisson relapse. The Poisson conveyance will display the likelihood of the occasion y as indicated by the Poisson interaction, specifically The Poisson distribution describes such situations more appropriately. So, we assume that the study variable x is a count variable and follows a Poisson distribution with parameter $\lambda > 0$ as

$$p(x) = \frac{e^{-\lambda}\lambda^{x}}{x!}, x = 0,1,2,\ldots \quad (1)$$

Note that the mean and variance of a Poisson random variable are same and related as

$$E(x) = \lambda, Var(x) = \lambda \quad (2)$$

.

Based on a sample $x_1, x_2, \ldots, x_n$, we can write $E(x_1) = \lambda$

and express the Poisson regression model as

$$x_i = E(x_1) + \varepsilon_1, i = 1,2,\ldots,n \quad (3)$$

.

Where ε_i's are disturbance terms. We can define a link function g that relates to the mean of study variable to a linear predictor as

$$g(\lambda_1) = \eta_i = \theta_0 + \theta_1 y_1 + \ldots + \theta_k y_k = y'_i \theta. \quad (4)$$

and

$$\lambda_i = g^{-1}(\eta_i) = g^{-1}(y'_i \theta) \quad (5)$$

.

The identity link function is $g(\lambda_i) = \lambda_i = y'_i \theta$. The log-link function is $g(\lambda_i) = ln\ ln\ \lambda_i = y'_i \theta$.

2.2. Poisson Regression Model

To estimate data in Poisson regression, we must observe the response variable model as a function of the predictor variable (Junadi, 1995). The value of y_i and x_i is related to the n dependent of variable y_i , where i is the observation, for example

$$\mu_i = e^{\beta_0 + \beta_1 x_i} \text{ and } y_i = \mu_i + \epsilon_i \text{ ,} \quad (6)$$

.

ϵ_i is the random variable, it can be written as

$$log\ (\mu_i) = \beta_0 + \beta_1 x_i. \quad (7)$$

the equation above is called log linear that relate between y and x.

Each of y_i has a Poisson distribution with the mean μ_i, the probability of y_i at the value of x_i is

$$P(y_i) = \frac{e^{\mu_i} \mu_i^{y_i}}{y_i!} = \frac{e^{(\beta_0 + \beta_1 x_i)} (\beta_0 + \beta_1 x_i)^{y_i}}{y_i!} \quad (8)$$

Poisson regression models express the mean of a discrete distribution as a function of the predictor variables (Junadi, 1995).

$$\mu_i = \exp\ (\boldsymbol{x}^T \boldsymbol{\beta}) \quad (9)$$

where

$$\boldsymbol{\beta} = [\beta_0\ \beta_1\ \beta_2 \ldots\ldots\ldots\ \boldsymbol{\beta_k}]^T, \boldsymbol{x} = [1\ x_{1i}\ x_{2i}\ x_{3i} \ldots\ldots\ldots\ x_{ki}]^T$$

After that, we need to estimate the parameters $\beta = (\beta_0, \ldots, \beta_k)$. In order to use Poisson regression, there are assumptions that must be met in the formation of the Poisson regression, the assumption is:

2.2.1. The Dependent Variable

In calculating Poisson regression following the experiment (discrete data), the frequency but not the measurement results.

2.2.2. Multicollinearity Test

In testing, the predictor variable must be free from multicollinearity. multicollinearity is the correlation between predictor variables. We must eliminate variables from the regression model so that the best model can be selected and avoid multicollinearity.

2.3. Parameter Estimation

To define the parameter estimated for Poisson regression, maximum likelihood estimation (MLE) is selected as a computation method. This method is maximizing the values of likelihood function in order to find the best parameter estimation.

Supposing n Poisson random variable yi, where $i = 1, 2, \ldots, n$ are taken independently, then the likelihood function of this distribution is:

$$L(y,\mu) = \prod_{i=1}^{n} f(y_i,\mu) = \prod_{i=1}^{n} \left\{\frac{\mu^{y_i} e^{-\mu}}{y_i!}\right\} = \frac{\left\{\prod_{i=1}^{n} exp\,(\boldsymbol{x}^T\boldsymbol{\beta})^{\sum_{i=1}^{n} y_i}\right\} exp\,(\boldsymbol{x}^T\boldsymbol{\beta})}{\prod_{i=1}^{n} y_i!} \quad (10)$$

To obtain the maximum value of log-likelihood function, differentiate the function towards each parameter $\beta_0, \beta_1, \ldots, \beta_k$, provided that each derivative value is equal to 0.

$$\frac{\partial ln\, L\,(y,\beta)}{\partial \beta_1} = 0, \frac{\partial ln\, L\,(y,\beta)}{\partial 0 \beta_2} = 0, \frac{\partial ln\, L\,(y,\beta)}{\partial \beta_3} = 0, \frac{\partial ln\, L\,(y,\beta)}{\partial \beta_k} = 0 \quad (11)$$

After determine the estimation of parameter $\beta_0, \beta_1, \ldots, \beta_k$, the estimation of Poisson Regression model can be written as:

$$\begin{aligned} \hat{y}_i &= \hat{\mu}_i + \varepsilon_i, \hat{y}_i = exp\left(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_j x_{ij}\right) + \varepsilon_i \\ &= \hat{\mu}_i + \varepsilon_i, \hat{y}_i = exp\left(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_j x_{ij}\right) + \varepsilon_i \end{aligned} \quad (12)$$

2.4. Best Model Selection

2.4.1. Goodness of Fit Test

In goodness of fit tests, deviance is used to measure the meaning of each coefficient and to test the good model used in Poisson model and logistic model cases. Deviance analysis can be implemented in all exponential distribution branches.

Deviance could be used for hypothesis testing, which is testing the value of parameters in Poisson regression. The contribution of each predictor variable can be calculated by how much it is decreasing the deviance value. Assume that

$$D\left(\beta_1, \beta_2, \ldots, \beta_{j-1}, \beta_{j+1}, \ldots, \beta_k\right) = D\left(\beta_1, \beta_2, \ldots, \beta_{j-1}, \beta_{j+1}, \ldots, \beta_k\right) - D(\beta_1, \ldots, \beta_k) \quad (13)$$

$D(\beta_1, \ldots, \beta_k)$ is devian calculated from all parameters in the model, while $D\left(\beta_1, \beta_2, \ldots, \beta_{j-1}, \beta_{j+1}, \ldots, \beta_k\right)$ is devian calculated without involving $\beta_j x_j$ in the model. The difference of devian value which does not involving $\beta_j x_j$ can be calculated by equation below:

$$D\left(\beta_1, \beta_2, \ldots, \beta_{j-1}, \beta_{j+1}, \ldots, \beta_k\right) = -2\, log\, log\left[\frac{L\left(\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_{j-1}, \hat{\beta}_{j+1}, \ldots, \hat{\beta}_k,\right)}{L\left(\hat{\beta}_1, \ldots, \hat{\beta}_k\right)}\right] \quad (14)$$

Those models show the difference of 2log L between complete model and reduction model. Statistic from ratio likelihood of $D\left(\beta_1, \beta_2, \ldots, \beta_{j-1}, \beta_{j+1}, \ldots, \beta_k\right)$ has Chi-Square distribution with q degree of freedom. The formula above could also be used to test each coefficient in the model, which is:

H0: $\beta_j = 0$ and H1: $\beta_j \neq 0$

Reject to H0 if $D\left(\beta_1, \beta_2, \ldots, \beta_{j-1}, \beta_{j+1}, \ldots, \beta_k\right) > X^2(a; 1)$

To determine a fit model and significant coefficient, the hypothesis testing is needed where type of test used is:

1. Joint test. To test which regression model is proper to use.
2. Coefficient test. To test whether each coefficient has an effect on the model.

2.4.2. Determination Coefficient of R^2

The determination coefficient (R^2) in linear regression analysis based on the use of sums-of-square with the least square method. It is widely used since it could describe a regression correlation between a dependent variable and predictor variable. The higher the value of R^2 ($0 < R^2 < 1$), the more accurate estimation obtained from the regression model. In regression model, log-likelihood is used to get the most

accurate analog of determination coefficient R^2 (Myers, 1990). Parameter estimation shows that if the parameter of Poisson regression model is $\beta_0, \beta_1, \beta_2, \beta_3, \dots, \beta_k$ and predictor variable $x, x_2, x_3, \dots, x_k$, then constant β_0 in the model given by devians of k degree of freedom as

$$D(\beta_1, \beta_2, \dots, \beta_k|\beta_0) = 2\ log\ log\ L(\beta_1, \beta_2, \dots, \beta_k)\ - 2\ log\ log\ L(\ \beta_0)$$
$$D(\beta_1, \beta_2, \dots, \beta_k) = 2\ log\ log\ L(\beta_0, \beta_1, \beta_2, \dots, \beta_k)\ , \quad (15)$$

while analog in sums-of-square is deviance from a model with constant β_0, then the determination coefficient R^2 as follow

$$R^2 = \frac{JK_{REG}}{JK_{TOTAL}} == \frac{2loglog\ L(\beta_1,\beta_2,\dots,\beta_k) - 2loglog\ L(\ \beta_0)}{2loglog\ L(\beta_1,\beta_2,\dots,\beta_k)} = 1 - \frac{loglog\ L(\beta_0)}{loglog\ L(\beta_0,\beta_1,\dots,\beta_k)}\ . \quad (16)$$

3. Research Method

3.1. Data Preparation

The data that we used is the number of cases of Dengue in 34 provinces of Indonesia, density population, and poverty level in 2015 from Badan Pusat Statistik (BPS, 2015).

3.2. Response and Predictor Variable

We have to determine which variable becomes an predictor variable and response variable. An predictor variable is a type of variable that has effects on other variables. A dependent variable is a type of variable that is influenced by the predictor variable. We determine the Case of dengue (Y) variable becomes our dependent variable. While density population (X1), poverty level (X2), and temperature (X3) become our predictor variable.

4. Analysis and Discussion

4.1. Multicollinearity test

In this multicollinearity test we used the VIF test, that the value must be more than 0.1 and less than 10. Because our data is more than 0.1 and less than 10, then the data passes the multicollinearity test.

Table 1. VIF test

X_1	X_2	X_3
1.071598	1.068367	1.003583

4.2. Poisson Regression Model

Here, we present Poisson Regression Model for dengue case in Indonesia in 2015. Parameter estimation for the model is given at Table 2.

Table 2. Parameter Estimation for Poisson Regression Model

	Estimate	Std. Error	Z value	$Pr\ (>\|z\|)$
(Intercept)	5.197	0.06824	76.15	$< 2 \times 10^{-16}$
X_1	3.181×10^{-5}	8.062×10^{-7}	39.46	$< 2 \times 10^{-16}$
X_2	-0.04830	0.0005717e-04	-84.48	$< 2 \times 10^{-16}$
X_3	0.09901	0.001913	51.76	$< 2 \times 10^{-16}$

In concurrent tests the Deviance must be more than Chi Square. In this data we used the 95% confidence level and 3 degree of freedom. We got the Deviance 13811 and Chi Square value 7.814728. Because Deviance value is more than Chi Square, then at least one predictor variable is significant. In a partial test the P-value must be less than 0.05, it means that predictor variable is significant. As we can see at Table 2, the P-value of X_1, X_2, and X_3 is less than 0.05, then all predictor variables are significant. According to

Table 2, Poisson Regression Model can be expressed below

$$\mu = exp\,(5.197 - 0.00003181X_1 - 0.04834X_2 + 0.099012X_3). \quad (17)$$

4.3. Overdispersion

In Poisson regression, the variance must be equal to the mean ($var = \mu$). Over dispersion is when the variance is more than mean. To check the over dispersion, we need to divide the Residual deviance and degree of freedom and if the value is more than 1 our data is called over dispersion. Table 3 provide the Null deviance and residual deviance from the Poisson Regression Model. Our data is over dispersion because the value is 5.405.

Table 3. Overdispersion

Null deviance:	175985 on 33 degrees freedom
Residual deviance:	162174 on 30 degrees freedom

4.4. Generalized Poisson Regression

We need to make the generalized Poisson regression into our data since it is over dispersion. We can see from the Table 4 that P-value of X_2 and X_3 less than 0.05 then our data significantly affect the predictor variable. The equation is:

$$\mu = exp\,(-0.9418 + 2.204 + 0.0000649X_2 - 0.04449X_2 + 0.267X_3). \quad (18)$$

Table 4. Parameter Estimation for Generalized Poisson Regression

	Estimate	Std. Error	Z value	$Pr\,(>\|z\|)$
(Intercept): 1	-0.9418	2.830	-0.333	0.7392254
(Intercept): 2	2.204	0.06385	34.521	$<2 \times 10^{-16}$
X_1	6.497×10^{-5}	3.965×10^{-5}	1.638	0.101336
X_2	-0.04449	0.02030	-2.191	0.028422
X_3	0.2670	0.07762	3.440	0.000583

4.5. AIC (Akaike Information Criterion)

In the AIC test, we need to see the smallest AIC value between the original Poisson Model and generalized Poisson Model. Table 5 shows that generalized Poisson Model have the smallest AIC than another one. Thus, we select the Generalized Poisson Regression as best model.

Table 5. AIC (Akaike Information Criterion)

	Poisson Regression Model	Generalized Poisson Model
AIC	162496.5	621.5827

5. Conclusion

In this paper, we have discussed the Poisson regression model and the generalized Poisson regression model to analyse the factors that have a significant effect on dengue cases in Indonesia in 2015. The best model is Generalized Poisson Regression because it has the smallest AIC. Based on the model, poverty and temperature have a significant contribution on the number of dengue 2015 cases in Indonesia. The interpretation for β_2 is every 1% increase in the percentage of poverty will cause a decrease in dengue cases to exp (-0.04449) = 0.9564 times from the initial dengue case assuming other predictor variables are constant, while The interpretation for β_3 is every 1^0 Celsius increase in temperature will cause an increase in dengue cases to exp (0.206) = 1,306 times from the initial dengue case assuming the other predictor

variables are constant. To help the government, the health sector, and the general public, hopefully, this research can give benefits for reducing dengue cases by paying attention to the factors that influence dengue cases.

References

Badan Pusat Statistik. (2015). *Kemiskinan, Kepadatan Penduduk, Kasus DBD 2015.* Jakarta: Badan Pusat Statistik

Sukohar A. 2014. *Demam berdarah dengue*. Medula. Bandar Lampung: Fakultas Kedokteran Universitas Lampung.2(2):1-14.

Widiyanto T. 2007. Kajian manajemen lingkungan terhadap kejadian demam berdarah dengue (DBD) di kota purwokerto jawa tengah. Semarang: Universitas Diponegoro.Page: 8-37.

Prasetyani, R.D. 2015. Faktor-Faktor yang Berhubungan dengan Kejadian Demam Berdarah Dengue. Bandar Lampung: Fakultas Kedokteran Universitas Lampung.

Desniawati F. 2014. Pelaksanaan 3M Plus Terhadap Keberadaan Larva Aedes Aegypti di Wilayah Kerja Puskesmas Ciputat Kota Tangerang Selatan Bulan Mei-Juni 2014. Jakarta: Universitas Islam Negeri Syarif Hidayatullah, Page: 8-38.

Marta Sundari, & Pardomuan Robinson Sihombing. (2021). PENANGANAN OVERDISPERSI PADA REGRESI POISSON: (Studi Kasus: Pengaruh Faktor Iklim Terhadap Jumlah Penderita Penyakit Demam Berdarah di Kota Bogor). Jurnal Lebesgue : Jurnal Ilmiah Pendidikan Matematika, Matematika Dan Statistika, 2(1), 1-9. https://doi.org/10.46306/lb.v2i1.48

Astuti, E. P., Fuadzy, H., & Prasetyowati, H. (2016). Pengaruh Kesehatan Lingkungan Pemukiman Terhadap Kejadian Demam Berdarah Dengue Berdasarkan Model Generalized Poisson Regression di Jawa Barat (Analisis Lanjut Riskesdas Tahun 2013). Bul Penelit Sist Kesehat. 2016; 19 (1): 109–117. Bul Penelit Sist Kesehat, 19(1), 109-117.

Pradhani, F. A. (2016). Perbandingan Model Neural Networks Dengan Poisson Regression Dan Negative Binomial Regression Pada Kasus Demam Berdarah Dengue (DBD) Di Surabaya (Doctoral dissertation, Institut Teknologi Sepuluh Nopember).

Junadi, P. (1995). Pengantar Analisis Data. Rineka Cipta. Jakarta