Journal of Mathematics Technology and Education **Vol. 1, No. 1, 2021 | 94-102**

Poisson Regresion Modeling Case Study Dengue Fever in Medan City in 2019

Juhenni Putri Sinaga[1], Ujian Sinulingga [2]

[1]Students of Mathematics, Departement of Mathematics, University of Sumatera Utara, Medan, 20155, Indonesia

[2]Departement of Mathematics, University of Sumatera Utara, Medan, 20155, Indonesia

Abstract. Dengue Hemorrhagic Fever (DHF) is an infectious disease caused by the dengue virus carried by the Aedes aegypti or Aedes albopictus mosquito which is spread in Southeast Asia. Medan City is one of the endemic areas for dengue fever in North Sumatra Province. This study aims to model the variable cases of dengue fever and determine the factors that have a significant effect on cases of dengue fever in the city of Medan. The method used in modeling the DHF case variable is the Poisson regression method with the response variable (Y) namely the number of DHF cases in Medan City, while the predictor variables are population density, number of health workers, number of health facilities, area height, and average waste production. In Poisson regression analysis, the response variable (Y) must meet the assumption of equidispersion. However, the assumption is often violated, namely overdispersion. Then Negative Binomial Regression was chosen as a non-linear model derived from the Poisson-gamma mixture distribution which is the application of the Generalized Linear Model (GLM) which describes the relationship between the response variable (Y) and the predictor variable (X).

Keyword: Dengue Fever, Equidispersion, Poisson Regresion, Negative Binomial Regresion, Overdispersion.

Abstrak. *Demam Berdarah Dengue (DBD) adalah penyakit infeksi yang disebabkan oleh virus dengue yang dibawa oleh nyamuk Aedes aegypti atau Aedes albopictus yang tersebar di wilayah Asia Tenggara. Kota Medan merupakan salah satu wilayah endemis demam berdarah di Provinsi Sumatera Utara. Penelitian ini bertujuan memodelkan variabel kasus demam berdarah dan mengetahui faktor yang berpengaruh signifikan pada kasus DBD di Kota Medan. Metode yang digunakan dalam memodelkan variabel kasus DBD adalah metode regresi Poisson dengan variabel respon (Y) yaitu jumlah kasus DBD Kota medan sedangkan variabel prediktornya yaitu kepadatan penduduk, jumlah tenaga kesehatan, jumlah sarana kesehatan, ketinggian wilayah, dan rata-rata produksi sampah. Dalam analisis regresi Poisson, variabel respon (Y) harus memenuhi asumsi equidispersion. Namun, seringkali terjadi pelanggaran asumsi yaitu terjadi overdispersion. Maka Regresi Binomial Negatif dipilih sebagai model non-linear yang berasal dari distribusi poisson-gamma mixture yang merupakan penerapan dari Generalized Linear Model (GLM) yang menggambarkan hubungan antara variabel respon (Y) dengan variabel prediktor (X).*

Kata Kunci: *Demam Berdarah, Equidispersi , Regresi Poisson, Regresi Binomial Negatif, Overdispersi.*

Received 08 Nov 2021 | Revised 11 Nov 2021 | Accepted 15 Nov 2021

*Corresponding author at: Departement of Mathematics, University of Sumatera Utara, Medan, 20155, Indonesia

E-mail address: juhenniputri@gmail.com, ujiansinulingga@usu.ac.id

e-ISSN: xxxx-xxxx | DOI: 10.32734/jomte.v1i1.7500
Journal Homepage: https://talenta.usu.ac.id/jomte

1. Introduction

Regression analysis is a method used to analyze the relationship between the response variable (Y) and several predictor variables (X). Regression analysis method which is generally used to analyze data with the response variable (Y) in the form of discrete data and predictor variable (X) in the form of discrete, continuous, categorical, mixed and count data with Poisson distribution is the Poisson regression model. Poisson regression model is derived from the Poisson distribution with parameter μ which depends on the predictor variable. In the Poisson regression model there is an assumption that must be met, namely equidispersion, which means the variance value of the Y variable must be equal to the average value.[1] In some cases it is often found that the observed data variance is greater than the average value, which is called overdispersion. Then the Negative Binomial Regression was chosen as a non-linear model derived from the Poisson-gamma mixture distribution which is the application of the Generalized Linear Model (GLM) which describes the relationship between the response variable (Y) and the predictor variable (X) and is used to model the data with the variable The response in the form of count data is used as an alternative to the Poisson regression model which has overdispersion, namely the variance value is greater than the average. The data used in this study is secondary data regarding the number of Dengue Fever in Medan City obtained from the Medan City Health Office and population density, number of health workers, number of health facilities, altitude of the area, and average waste production in Medan City, Central Bureau of Statistics of Sumatra. North. [2] In this study, there are problems, namely what are the factors that significantly influence the number of DHF cases in Medan City using Poisson regression and how to model the number of DHF cases in Medan City using Poisson regression.

2. Literature

2.1. Poisson Regression Model

Poisson regression is a regression model that can be used on data whose response variables are not normally distributed and of discrete type, namely Poisson distribution as the main requirement. Poisson regression is an application of the Generalized Linear Model (GLM) which describes the relationship between the response variable (Y) in the form of discrete data or count data and predictor variable (X) in the form of discrete, continuous, categorical or mixed data. If the response variable (Y) is discrete data with a Poisson distribution with parameters $\mu > 0$, where μ is the average of the response variables (Y), then the probability function is [5]:

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

In GLM there is a linear function and relates the average value of the response variable with an explanatory variable (predictor), namely [4]

$$g(\mu_i) = \eta_i = x_i^T \beta \tag{2}$$

$$\mu_i = \beta_0 + \beta_1 X_{i_1} + \beta_1 X_{i_2} + \ldots + \beta_p X_{i_p} + \varepsilon_i \tag{3}$$

The function g is called the link function. The relationship between the mean value and the linear explanatory variable is:

$$\mu_i = g^{(-1)}(\beta_i) = g^{(-1)}(x_i^T \beta) \tag{4}$$

In the Poisson regression model, usually the connecting function is a log linking function, because the average of the response variables will be in the form of an exponential function and guarantee that the value of the estimated variable of the response variable will be non-negative. The log linking function is of the form:

$$g(\mu_i) = ln\mu_i = x_i^T \beta \tag{5}$$

The relationship between the average value of the response variable and the linear explanatory variable is as follows:

$$\ln[(\mu_i)] = x_i^T \beta \tag{6}$$

We take the exponential function on both sides, we get:

$$e^{([\ln(\mu_i)]} = e^{(x_i^T \beta)} \tag{7}$$

$$\mu_i = e^{(x_i^T \beta)} \tag{8}$$

So that the connecting function for the multiple Poisson regression model can be written as follows (Rini Cahyandari, 2014):

$$\mu_i = e^{(x_i^T \beta)} \tag{9}$$

$$\mu_i = e^{(\beta_0 + \beta_1 X_{i_1} + \beta_2 X_{i_2} + \ldots + \beta_p X_{i_p} + \varepsilon_i)} \tag{10}$$

Then the Poisson Regression model is obtained as follows:

$$Y = e^{(\beta_0 + \beta_1 X_{i_1} + \beta_2 X_{i_2} + \ldots + \beta_p X_{i_p} + \varepsilon_i)} \tag{11}$$

2.2. Negative Binomial Regression Model

Negative Binomial Regression (RBN) has a Negative Binomial distribution to model the count data experiencing equidispersion or overdispersion. RBN does not have a requirement that the variance value must be equal to the average value, so this model is chosen as an alternative that

will get the best model. RBN uses the GLM model with a log link function which has the same equation form as the Poisson regression equation, namely:[6]

$$Y = e^{(x_i^T \beta)} \tag{12}$$

2.3. Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov test is a goodness of fit test, meaning that what is considered is the level of conformity between the distribution of a series of observed sample values with a certain theoretical distribution.

2.4. Multicollinearity

The term multiple collinearity multicollinearity was first coined by Ragnar Frish which means that there is a perfect or exact linear relationship between the independent variables or predictors in the regression model. The detection of multicollinearity cases was carried out using the Variance Inflation Factor (VIF) value criterion. If the VIF value is greater than 10, it indicates the presence of multicollinearity between predictor variables. [3]

2.5. Parameter Significance Test for Poisson Regression and Negative Binomial Regression

The significance test of the model is needed to see the effect of the response variables scattered in the model. The significance test of the model is distinguished into the Overall test (simultaneously) and the Partial test [6].

2.6. Best Regression Model Selection

Comparing the Poisson regression model and the negative Binomial regression model is to get the best modeling on the response variable or on the variable number of cases of DHF in Medan City. In selecting the best model by looking at the AIC value (*Akaike Information Criteria*).

3. Methodology

3.1. Data Source

The data used in this study is secondary data regarding the number of dengue cases in Medan City, population density, number of health workers, number of health facilities, altitude of the area, and average waste production. The data used was obtained from the publication of the City of Medan in Figures in 2019 by the Central Statistics Agency of North Sumatra Province and the Medan City Health Office. The data are data for each of the 21 sub-districts in Medan City.

3.2. Data Analysis Techniques

Data analysis in this study using SPSS software. As for the steps of data analysis to be carried out are as follows:

1. Describe the characteristics of each data variable taken
2. Calculate the mean and variance of each data variable.
3. Kolmogorov-Smirnov test
4. Multicollinearity testing
5. Modeling Poisson Regression on the number of DHF cases in Medan City
   (a) Testing the significance of the Poisson regression model simultaneously using the Deviance test
   (b) Testing the significance of the partial Poisson regression model using the Wald Chi-Square test.

4. Results and Discussions

4.1. Descriptive Statistics

Based on the calculations in this study, the average value = 50.86 and the variance value = 805.329 because the variance value is greater than the average value, it indicates that the data is overdispersion.

4.2. Kolmogorov-Smirnov Test

Kolmogorov-Smirnov test to determine whether the response variable (Y) has a Poisson distribution. Hypothetical steps as follows: Hypothesis
H_0 : response variable data with poisson distribution
H_1 : response variable data is not distributed Poisson
Significance Level
$\alpha = 0.05$
Statistics Test
Reject H_0 if $D > D_{(\alpha;n)}$, where $D_{(\alpha;n)}$ is a critical value obtained from the Kolmogorov-Smirnov table or using reject H_0 if $p_(value) \leq \alpha$.
By using statistical test steps obtained a value of $D = 0,188 < D_{(\alpha;21)=0,287}$ or $p_(value) = 0,194 > \alpha = 0,05$ then H_0 is accepted. So it can be concluded that the response variable (Y) is Poisson distributed data.

4.3. Multicollinearity Test

Multicollinearity testing is used to determine whether or not there is a deviation from the classical assumption, namely whether there is a relationship between predictor variables in the regression model.

Table 1. Multicollinearity Test.

Variable	Tollerance	VIF	Decision
X_1X_2	0.95	1.0526	No multicollinearity
X_1X_3	.9239	1.0823	No multicollinearity
${}_1X_4$	0.9967	1.0033	No multicollinearity
X_1X_5	0.9999	1.0000	No multicollinearity
X_2X_3	0.9643	1.0370	No multicollinearity
X_2X_4	0.8745	1.1435	No multicollinearity
X_2X_5	0.9319	1.0731	No multicollinearity
X_3X_4	0.9987	1.0013	No multicollinearity
X_3X_5	0.9481	1.0547	No multicollinearity
X_4X_5	0,8876	1.1266	No multicollinearity

In table 1 it is concluded that there is no multicollinearity in the observed data, which means that there is no relationship between the predictor variables. Then the data analysis using the Poisson regression model can be continued.

4.4. Poisson Regression Model

Modeling the Poisson regression first looks for the estimated value of the parameter β_0, β_1, β_2, β_3, β_4, β_5. The parameter estimate β was searched using SPSS. Then the result of parameter estimation β is as follows:

Table 2. Estimated value of the Poisson regression parameter.

Parameters	Estimates β	Standard Error
β_0	(Constant)2.134	0.2438
β_1	-0.0004463	0.000007406
β_2	0.005	0.0015
β_3	0.017	0.0106
β_4	0.011	0.0028
β_5	0.016	0.0013

Based on table 2, substitute the values into the Poisson regression equation as follows:

$$Y = e^{(\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3+\beta_4X_4+\beta_5X_5)}$$

$$Y = e^{(2.134-0.00004463X_1+0.005X_2+0.017X_3+0.011X_4+0.016X_5)}$$

4.5. Overdispersion

The Poisson regression overdispersion test can be seen from the deviance and Pearson chi square values for the response variable (Y). Then the following results are obtained:

Table 3. Poisson Regression Overdispersion Test.

Test Name	Value	Value/db	$\chi^2_{(0.05;15)}$	Decision
Deviance	85,223	5,682	24,996	Overdispersion
Pearson chi-square	78,363	5,224	24,996	Overdispersion

Based on the calculation results above, it can be seen that reject H_0 because *Deviance* = 85,223 > $\chi^2_{(0,05;15)}$ = 24,99 and Pearson chi square = 78,363 > $\chi^2_{(0.05;15)}$ = 24.99 , it can be concluded that there is overdispersion in the response variable (Y).

4.6. Negative Binomial Regression Model

The negative binomial regression overdispersion test was conducted to test whether there was an overdisperion in the response variable (Y) in which the RBN had a fairly large overdisperion. Then the following results are obtained:

Table 4. Results of the Negative Binomial Regression Model.

Variable	Value β	Value W (Wald)	Value $(\chi^2)_{(0.05;1)}$	Decision
Constant	2,270	2,3691	3,841	Not Significant
X_1	-0.0003837	0.6967	3.841	Not Significant
X_2	0.002	0.0104	3.841	Not Significant
X_3	0.006	0.0587	3.841	Not Significant
X_4	0.011	0.2744	3.841	Not Significant
X_5	0.018	4.0904	3.841	Significant

Based on the results of the calculations above. So the Negative Binomial Regression Model:

$$Y = e^{(\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3+\beta_4X_4+\beta_5X_5)}$$

$$Y = e^{(2.270-0.00003837X_1+0.002X_2+0.006X_3+0.011X_4+0.018X_5)}$$

It can be seen that for the predictor variable (X_5) the value of W=4,0904 > $\chi^2_{(0.05;15)}$ = 3.841, so the parameter β_5 significant effect on the level of 0.05 which means the average waste production factor (X_5) has a significant effect on the number of cases of DHF (Y).

4.7. Overdispersion

The Poisson regression overdispersion test can be seen from the Deviance and Pearson chi square values for the response variable (Y). Then the following results are obtained:

Table 5. Negative Binomial Regression Overdisperion Test.

Test Name	Value	Db	Value/df	$\chi^2_{(0.05;15)}$	Decision
Deviance	1.952	15	0.130	24,996	No overdispersion
Pearson chi-square	1.561	15	0.104	24,996	No overdispersion

Based on the calculation results above, it can be seen that receiving H_0 because *Deviance* = 1.952 $< \chi^2_{(0.05;15)}$ = 24,996 then it can be concluded that there is no overdispersion in the response variable (Y) and also a decrease in the level of overdispersion when using the Negative Binomial regression model.

4.8. Best Regression Model Selection

Comparing the Poisson regression model and the negative Binomial regression model is to get the best modeling on the response variable or on the variable number of cases of DHF in Medan City. In selecting the best model by looking at the AIC value (*Akaike Information Criteria*) which has the smallest value.

Table 6. Deviance and Pearson chi-square values.

Value	Poisson Regression	Negative Binomial Regression
AIC	214,801	214,308

In the table above, it can be seen that the AIC value in the Negative Binomial regression model is much smaller than the Poisson regression model. So it can be concluded that the best model between the Poisson regression model and the Negative Binomial regression is the Negative Binomial regression model.

5. Conclusions

Based on the above calculations, some conclusions can be drawn as follows: The factors that influence the number of DHF cases (Y) with calculations using the Poisson regression model are population density (X_1), number of health workers (X_2), area height (X_4), and average waste production (X_5) has a significant effect. And the factor that does not affect the number of DHF cases is the number of health facilities (X_3). The Poisson regression model obtained from the calculation is

$$Y = e^{(2.134-0.00004463X_1+0.005X_2+0.017X_3+0.011X_4+0.016X_5)}$$

The factors that affect the number of DHF cases (Y) in the calculation using the Negative Binomial regression model are the average waste production (X_5). The negative binomial regression model obtained from the calculation is

$$Y = e^{(2.270-0.00003837X_1+0.002X_2+0.006X_3+0.011X_4+0.018X_5)}$$

REFERENCES

[1] Cameron Ca, Trivedi PK. *Regression Analysis of Count Data*. Cambridge : Cambridge University Press. 1998.

[2] Yoeyoen A Indrayani, Tri Wahyudi. Situas Penyakit Demam Berdarah di Indonesia tahun 2017. *Info Datin Pusat Data dan Indormasi Kesehatan RI*. 2018.

[3] Algifari. *Analisis Regresi Teori, Kasus, dan Solusi*. Yogyakarta: BPFE-Yogyakarta. 2000.

[4] Rini C. Pengujian Overdispersi pada Model Regresi Poisson (Studi Kasus Laka Lantas Mobil Penumpang di Provinsi Jawa Barat). *Jurnal Statistika*, Vol. 14 No. 2, 69 - 76. 2014.

[5] Rahmadeni, Zulya D. Perbandingan Model Regresi Generalized Poisson dan Binomial Negatif Untuk Mengatasi Overdispersi Pada Regresi Poisson. *Jurnal Sains Matematika FMIPA Universitas Udayana, Bukit Jimbaran - Bali*. Vol. 2 No. 2. 2016.

[6] Rara KNM, Wayan IS S, Luh NPS. Perbandingan Regresi Binomial Negatif dan Regresi Generalisasi Poisson dalam Mengatasi Overdispersi.(Studi Kasus: Jumlah Tenaga Kerja Usaha Pencetak Genteng di Br. Dukuh, Desa Pejaten. *E-Jurnal Matematika Universitas Islam Bandung*,vol. 3, 107-115. 2014.