Proceedings of Science and Mathematics
Faculty of Science,
Universiti Teknologi Malaysia
Vol. 8, 2022, page 12-18

Generalized Linear Model Approach on Dengue Epidemic in Selangor

Nur A'ina Farahah Che Abdullah, Ong Chee Tiong
Department of Mathematical Sciences, Faculty of Science
Universiti Teknologi Malaysia, 81310 Johor Bahru, Malaysia

Abstract

Dengue is an arthropod-borne viral disease that carried by a female mosquito called Aedes aegypti. Dengue viruses (DENV) are a part of a genus of viruses in the family Flaviviridae. They are recognized by their four serologically and genetically unique serotypes that could cause dengue. Because of its rapid spreading and serious effects that could lead to loss of life, it has become a worried situation particularly for Malaysian health authorities. Changes in climate factors are ordinary processes that may significantly affect the occurrence of the transmittable diseases like dengue. In this study, the climate factors such as temperature, rainfall, humidity and wind speed were considered. Some statistical models were applied to recognize the climatic factors associated with transmittable illnesses. Since the interactions between the climate variables and the disease are complex and nonlinear, demonstrating their interactions have become the core challenge in climate-health studies. Therefore, a statistical approach called Generalized Linear Model (GLM) with Poisson and Negative Binomial has been proposed in this study to observe the effects of the climate factors on dengue incidence by considering the collinearity between the variables. This study only focuses in the state of Selangor, being the dengue hotspot in Malaysia and the data collected is only in the year 2017. There was moderate collinearity detected between the climate factors. The climate variables which is temperature and rainfall were statistically significant with dengue cases, while humidity and wind speed were insignificant. Having this kind of relationships, this model could help to regulate and prevent the spread and transmittable diseases like dengue can be avoided for a better lifestyle.

Keywords: Dengue cases; Linear model; Generalized linear model; Poisson GLM; Negative-binomial GLM; Count data

1. Introduction

Patients with dengue fever may experience high fever, headache, vomiting, muscle and joint pain, and a characteristic rash. An estimated 2.5 billion people in tropical and subtropical countries are at risk [1]. Dengue fever has become a significant priority for global health organisations due to its rapid spread. Currently, forty percent of the world's population is at danger of contracting the disease [2]. The average annual cost of dengue illness in the Americas is $2.1 billion [3].

In Malaysia, dengue fever epidemics have been observed for a very long period. The number of cases has continued to climb despite the Ministry of Health's greatest efforts to manage the disease. Even the local temperature variations associated with the rapid increase in dengue cases are poorly understood.

The aims for this study are to (1) identify the relationship between climate factors and dengue disease, (2) analyse the effect of climate parameters on dengue disease by applying the generalized linear model and (3) investigate the strength of the relationship between two variables by referring to the correlation coefficient value.

2. Literature Review

2.1. Dengue in Malaysia

Malaysia with 27.7 million population, has consistently documented for having a growing dengue infection each year since 1980. The very first dengue incidence reported in Malaysia was from Penang state, in December 1901, following reports of cases in Hong Kong, Bangkok, and Singapore within the same year. There were major nationwide dengue epidemics recorded in a four-year cycle, which in 1974, 1978, 1982 and 1990 [4]. A Health Fact that was provided by Ministry of Health Malaysia in 2008, the incidence rate of dengue recorded was 167.76 per 100 000 people, with 0.02 fatality rate.

Recent studies suggest that weather affects the life cycle, activity, biting rates, and incubation periods of vectors, which affects the severity of epidemics in Malaysia [5]. Weather conditions affect Aedes mosquito life cycles. Aedes albopictus is being found in the Northern Hemisphere, where it is not endemic. 50–60% of the world's population will be exposed to these vectors in 100 years, compared to 35% now [6].

2.2. Dengue and Climate Change

In the 21st century, climate change is recognized as the greatest health threat [7]. Previous research on the future spread of Aedes mosquitoes in Australia are based on potential future climate change scenarios and some researches have projected that their spreading may be to the whole nation. In addition to influencing the dynamics of dengue transmission, climate change influences the habits and lives of people.

There was a study that applied the multiple linear regression on the temperature and rainfall factors to explore their effect on dengue incidence in Metro Manila. By using the data from 1996 to 2005, it showed that dengue was only influenced by the changing pattern of rainfall [8].

2.3. Generalized Linear Model

The generalized linear model (GLM) is most frequently employed in dengue research. GLM may accept more complex data structures, including Gamma, Poisson, binomial, and other non-normal distributions. For instance, a study in Delhi used this model to examine seasonal variation and establish a probabilistic model of dengue predictors by using 2015-2018 dengue and environment data [9].

3. Methodology

3.1. Study Area

This investigation was conducted in Selangor, in the west of Peninsular Malaysia, which has a total area of 7951 square kilometers. It was separated into 9 districts which are Sabak Bernam, Kuala Selangor, Hulu Selangor, Gombak, Petaling, Klang, Sepang, Hulu Langat and Kuala Langat. Climate in Selangor is equatorial and heavily influenced by monsoons. There are two wet seasons: The Southwest monsoon and the Northeast monsoon. Selangor had a population of 6524.6 thousand, comprised of 3378.4 thousand men and 3146.3 thousand women.

3.2. Data Collection

Dengue hotspots data were gathered from an open source website called data MAMPU which were accessed from *data.gov.my*. This study only considered the dengue data in the year 2017 because the latest data that could be obtained was in 2017. Similarly, climatological elements, which referred to temperature, rainfall, humidity, and wind speed, were obtained from Visual Crossing website, which could be accessed from *https://www.visualcrossing.com/*. Both of the data were complete, provided by weekly and there was no missing value for that year of study. The climatological factors were set as independent variables (x). Meanwhile, dengue incidence data was selected as dependent variable (y). In this study, we only focused on the weekly data in the year of 2017 for both data (dengue and climate). Table 1 shows the variables that are measured in this study.

Table 1: List of Variables

Variables	**Description**
Y	Dengue cases
X_1	Mean rainfall (mm)
X_2	Mean temperature (℃)
X_3	Mean wind speed (km/h)
X_4	Mean relative humidity (%)

3.4 Statistical Methodology

3.4.1 Multicollinearity Detection

Collinearity diagnostic will focus on Chi-square value and determinant outcomes. To determine the collinearity between the explanatory variables, individual multicollinearity diagnostic measures will be applied. This study is limited to Variance Inflation Factor (VIF). The more collinearity between regressors means the greater VIF value. The VIF formula is given as equation (1).

$$VIF_j = \frac{1}{1-R_j^2}\ , \tag{1}$$

where:
j = the explanatory variable,
R_j^2 = the unmodified determination coefficient for regressing the remaining j-th variables and denotes as equation (2).

$$R_0^2 = R_{yx_1}^2 + R_{yx_2}^2 + \ \dots \ + R_{yx_p}^2. \tag{2}$$

3.4.2 Generalized Linear Model

The Poisson GLM will be applied since the dengue incidence is a counted data. Similarly, this approach is used when the response variables are non-normally distributed and only have positive integers. The distribution for the response variable Y is assumed to be in Poisson family. By referring to [10], the Poisson probability distribution with parameter λ is defined as equation (3).

$$f(Y_i;\lambda_i) \ = \ \frac{e^{-\lambda_i}\lambda^{Y_i}}{Y_i!}\ ,\ Y_i = 0,1,2,3,\dots,p\ , \tag{3}$$

where:
Y_i = random variables that represent the dengue cases in week i in a period p,
λ_i = the mean and variance of Y.

Let mean of the response variable, Y is given as equations (4) and (5).

$$\mu_i = E(Y_i) = \lambda_i\ , \tag{4}$$
$$\mu_i = exp({x_i}^T\beta) = exp(\eta_i). \tag{5}$$

where:
x_i^T = the transposed vector of the design matrix $\boldsymbol{X}$ of the $\imath$th row,
β = the estimated parameter.

Linear predictor [11] is given as equation (6):

$$\eta_i = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p\ \ , i = 1,2,\dots,p\ . \tag{6}$$

The mean of the response, μ_i and linear predictor, η_i is specified by a link function, g as in equation (7).

$$g(\ \mu_i) = \eta_i = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p\ , \tag{7}$$

where:
$\mu_i =$ a function of some independent variables via a log link function as below.

The log link function is defined as equation (8).

$$ln(\mu_i) = x_i{}^T\beta. \qquad (8)$$

The assumption in performing GLM is that mean and variance are equal [12]. A negative binomial regression is a generalization of Poisson regression, that can loosen the restrictive assumption that the variance is equal to the Poisson model's mean. It is used to deal with the data where over-dispersion exists. The expected value of the response is given as equation (4) while the variance is given as equation (8).

$$Var(y_i) = \mu_i + \frac{\mu_i{}^2}{\theta}. \qquad (8)$$

Suppose $\alpha = \frac{1}{\theta}$, then the variance function can be written as equation (9).

$$Var(y_i) = \mu_i + \frac{\mu_i{}^2}{\theta} = \mu_i + \alpha\mu_i{}^2 \ , \qquad (9)$$

where:
$\alpha =$ dispersion parameter.

As a result, the density function Y_i is can be defined as equation (10).

$$f(Y = y_i; \mu_i, \alpha^{-1}) = \frac{\Gamma(y_i+\alpha^{-1})}{\Gamma(y_i+1)\Gamma(\alpha^{-1})}\left(\frac{1}{1+\alpha\mu_i}\right)^{\alpha^{-1}}\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right)^{y_i}, \qquad (10)$$

where:
$\Gamma =$ the gamma integral specialized to a factorial integer argument [13].

For the GLM, the deviance takes the form of the likelihood ratio, and be defined as equation (11).

$$D(y,\mu) \equiv 2\sum_{i=1}^{n}\left[y_i \ ln\left(\frac{y_i}{\mu_i}\right) - (y_i + \alpha^{-1})\left(\frac{1+\alpha y_i}{1+\alpha\mu_i}\right)\right]. \qquad (11)$$

For the models that are estimated by the maximum likelihood, one way in comparing the non-nested model is through the Akaike Information Criterion (AIC), which is based on the fitted log-likelihood function as in equation (12).

$$AIC = -2D(y,\mu) + 2k. \qquad (12)$$

4. Results and discussion

4.1. Descriptive Statistics

Table 2 described the statistics of each variables that are being measured. The skewness of the dengue data is not equal to zero, which means that it is not normally distributed. Hence, the GLM is applied to deal with the data that consist of only non-negative integers and that is not normally distributed. Furthermore, the variance of the dengue is greater than the mean, which implies that there exists overdispersion in the data since the mean is not equal to the variance. This does not align with the assumption made. Then further analysis should be applied to deal with over disperse data.

Table 2: Distribution of Dengue Cases and Selected Weather Parameters in Sultan Abdul Aziz Shah Station in 2017

Variables	Mean	Standard Deviation	Variance	Skewness
Weekly dengue cases	870.96	262.3479	68826.4299	-0.1668

Variables	**Mean**	**Standard Deviation**	**Variance**	**Skewness**
Cumulative weekly rainfall (mm)	48.83	29.9933	89.5955	0.6066
Weekly mean temperature (℃)	27.90	0.7431	0.5523	-0.0298
Weekly mean wind speed (km/h)	16.23	1.4004	1.9611	0.4123
Mean relative humidity (%)	77.85	4.1090	16.9941	-0.3700

4.2. Model of Dengue Disease

4.2.1. Multicollinearity testing

Variance inflation factors were used to analyse the collinearity. VIF could identify the correlation between the independent variables and the strength of that correlation. The VIF value equals to 1 indicates that there is no correlation between the independent variables. The value between 1 to 5 suggest that there is a moderate correlation, but it is not severe enough to warrant the corrective measures. The VIF value greater than 5 represent the critical level of multicollinearity where the coefficients are poorly estimated, and the *p*-values are questionable. Table 3 showed the tolerance and VIF value for the independent variables. From the output that was obtained, there were only moderate correlation between the independent variables since all the VIF values were less than 5, so the collinearity was poor and could be ignored.

Table 3: Collinearity Diagnosis of The Independent Variables

Variables	**Tolerance**	**VIF**	**Detection**
Rainfall, X_1	0.4879	2.0496	Moderate collinear
Temperature, X_2	0.2603	3.8411	Moderate collinear
Wind speed, X_3	0.8983	1.1132	Moderate collinear
Humidity, X_4	0.2850	3.5082	Moderate collinear

4.2.2. Poisson and Negative Binomial of Generalized Linear Model

The symptoms of dengue fever are like sudden high fever or prolonged fever, severe joint or muscle paint, skin rashes and nausea. It usually takes three to seven days for the symptoms to appear after the infection, while for the skin rashes, it could occur in two to five days. So, in the study of dengue fever, a delay of one week is relevant and could be taken into consideration according to WHO. Hence, the model was based on the time lapse of one week and was applied throughout the analysis. In order to examine the significance of the relationship between the dengue disease and the weather conditions, the generalized linear model approach was used together with Poisson and Negative Binomial.

Table 4 displayed the AIC values for both Poisson and Negative Binomial. The AIC value for Poisson was 3742, which was relatively high compared to when using Negative Binomial which gave the AIC value of 727.63. The value was smaller when using Negative Binomial indicated that the Negative Binomial is much better than Poisson GLM when dealing with overdispersion. Hence, GLM with Negative Binomial will be used in the analysis as equation (13).

$$ln(Y) = -0.1560 + 0.0344X_1 + 0.2449X_2 + 0.0120X_3 - 0.0045X_4\,. \qquad (13)$$

The summary of analysis by using Negative Binomial was recorded in Table 5. The result showed that only rainfall and temperature that were statistically significant since the p-value were less than a 5% significant level, while the wind speed and humidity were not significant. This meant that the dengue cases were influenced by the rainfall and temperature variables, but not by the wind speed and humidity factors. It was not consistent with the early assumption that all the climate variables might influence the dengue cases in Selangor. Since p-value is more significant and should be prioritised compared to the VIF value, hence, the final model equation could be written as equation (14).

$$ln(Y) = -0.1560 + 0.0344X_1 + 0.2449X_2. \qquad (14)$$

The optimal model, including rainfall and temperature yielded a positive relationship with dengue cases. This indicated that the high amount of rainfall and the high temperature contributing to the increases in dengue cases in Selangor.

Table 4: The Result of AIC Value for One-week Time Lapse

Poisson GLM	Negative Binomial GLM
3742.0	727.63

Table 5: The Summary of The Analysis by Using Negative Binomial GLM

Coefficients	Estimated Value	Standard Error	*Z*-value	*p*-value
Intercept	-0.158723	3.986244	-0.040	0.96824
Rainfall, X_1	0.004918	0.001875	2.623	0.00872**
Temperature, X_2	0.245000	0.103618	2.364	0.01806*
Wind speed, X_3	0.011974	0.029594	0.405	0.68575
Humidity, X_4	-0.004529	0.017904	-0.253	0.80031

*** Most statistically significant with p-value < 0.05 in the range $(0,0.001]$

** Moderate statistically significant with p-value < 0.05 in the range $(0.001,0.01]$

* Less statistically significant with p-value < 0.05 in the range$(0.01,0.05]$

Conclusion

Based on this study, moderate collinearity was detected using the VIF value. Then the model could be generated by using the Negative Binomial for the further analysis. By looking at the p-value, the final model gave that the factors of rainfall and temperature could be used to predict the dengue cases in Selangor. It can be concluded that both variables are the vital factors that should be considered when designing the model for dengue.

Acknowledgement

The researcher would like to thank all people who have supported the research and contributed towards the understanding and thoughts.

References

[1] Halstead, S. B. (1988). Pathogenesis of dengue: Challenges to molecular biology. *Science (New York, N.Y.)*, *239*(4839), 476–481. https://doi.org/10.1126/science.3277268

[2] World Health Organization. (2009). Dengue Guidelines for Diagnosis, Treatment, Prevention, and Control. Special Programme for Research and Training in Tropical Diseases. World Health Organization, Geneva. https://www.who.int/tdr/publications/documents/dengue-diagnosis.pd

[3] Shepard, D. S., Coudeville, L., Halasa, Y. A., Zambrano, B., & Dayan, G. H. (2011). Economic impact of dengue illness in the Americas. *The American journal of tropical medicine and hygiene*, *84*(2), 200–207. https://doi.org/10.4269/ajtmh.2011.10-0503

[4] Lam, S. K. (1993). Two decades of dengue in Malaysia. *Tropical medicine*, *35*(4), 195-200.

[5] Hii, Y. L., Zaki, R. A., Aghamohammadi, N., & Rocklöv, J. (2016). Research on climate and dengue in Malaysia: A systematic review. *Current environmental health reports*, *3*(1), 81-90.

[6] Murray, N. E. A., Quam, M. B., & Wilder-Smith, A. (2013). Epidemiology of dengue: Past, present and future prospects. *Clinical epidemiology*, *5*, 299-309.

[7] Akter, R., Hu, W., Naish, S., Banu, S., & Tong, S. (2017). Joint effects of climate variability and socioecological factors on dengue transmission: Epidemiological evidence. *Tropical medicine & international health*, *22*(6), 656–669. https://doi.org/10.1111/tmi.12868

[8] Su, G. L. S. (2008). Correlation of climatic factors and dengue incidence in Metro Manila, Philippines. *AMBIO: A journal of the human environment*, *37*(4), 292-294. https://doi.org/10.1579/0044-7447(2008)37[292:COCFAD]2.0.CO;2

[9] Singh, P. S., & Chaturvedi, H. K. (2022). A retrospective study of environmental predictors of dengue in Delhi from 2015 to 2018 using the generalized linear model. *Scientific reports*, *12*(1), 1-10.

[10] Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the royal statistical society*, *135*(3), 370–384. https://doi.org/10.2307/2344614

[11] Dobson, A. J., & Barnett, A. G. (2018). *An introduction to generalized linear models* (4th ed., pp.49-50). A Chapman & Hall book/CRC press.

[12] Lawless, J. F. (1987). Negative binomial and mixed Poisson regression. *The Canadian journal of statistics, 15*(3), 209-225. https://doi.org/10.2307/3314912

[13] Breslow, N. E. (1984). Extra-Poisson variation in log-linear models. *Journal of the royal statistical society*, *33*(1), 38–44. https://doi.org/10.2307/2347661.