RESEARCH ARTICLE   **OPEN ACCESS**

DENGUE IN BRAZIL: REGRESSION AND PROBABILISTIC MODELING

Mickaelle Maria de Almeida Pereira, Felipe Fernando Ângelo Barrêto, Jucarlos Rufino de Freitas, *Josimar Mendes de Vasconcelos, Antonio Samuel Alves da Silva, Guilherme Rocha Moreira, Frank Gomes-Silva and Moacyr Cunha Filho

Department of Statistics and Informatics, Federal Rural University of Pernambuco, Recife - PE, Brazil

ABSTRACT

The work aimed to adjust Poisson models and their generalizations (Quasi-Poisson and Negative Binomial) to the behavior of the number of dengue cases grouped in Recife-PE. Weekly epidemiological records of dengue were used, from 2009 to 2018, and they were made available through the Citizen Information Service. Initially, Poisson models and their generalizations were applied, followed by a diagnostic analysis of the residuals, using influence measures, in order to analyze how much a specific observation affects some property of the studied model. The adjustment results indicated that the Negative Binomial model showed a satisfactory adjustment, compared to other models for description of the data, minimizing dispersions parameters more accurately. Regarding the analysis of the residuals, it can be stated that the assumptions of normal distribution and homogeneity of the residuals in the Generalized Linear Models (GLM) were provided. In addition, the existence of atypical and influential values was verified, which should be examined carefully. Therefore, it can be concluded that diagnosis is an important tool for researchers, in which they present basic premises for the reliability of the GLM, data set, and goodness of fit. Finally, we point out that more scientific studies are needed, in this context, to have more consistency in this application.

INTRODUCTION

Dengue is an arbovirus that causes great impacts on public health. According to a survey produced by the World Health Organization (WHO), it is estimated that between 100 and 400 million infections occur annually, caused by the Aedes aegypti mosquito (WHO, 2020). Brazil, being a tropical country and endemic to dengue, has conditions that favor the development and proliferation of the main transmitter of the disease (WHO, 2012; Oliveira *et al*., 2018; Sobral and Sobral, 2019). Conditions such as high rainfall, high temperatures, high relative humidity, in addition to social aspects such as urban peripheries and low income are related to a higher probability of contracting the disease (Bhatt *et al*., 2013; Ferreira *et al*., 2018; Sobral and Sobral, 2019). In Brazil, from 2010 to 2019, approximately 10 million cases of dengue were reported, with the Southeast and Northeast being the most affected Brazilian regions (Nunes *et al*., 2019; do Carmo *et al*., 2020). Analyzing a more specific interval for the Northeast (2013 to 2018), more than one million probable cases of dengue were registered (18.1% of the total cases in the country) and 707 deaths (22.3%). The state of Pernambuco obtained about 20.1% of the total probable cases and 19.4% of deaths in the Brazilian Northeast (do Nascimento *et al*., 2020). According to the epidemiological bulletin of the Department of Health Surveillance, there was an increase of 19.5% in the incidence of Dengue, in the state of Pernambuco, until Epidemiological Week (EW) 11 of 2019 compared to 2018 (Freitas *et al*., 2020). In the literature, there are several methodologies that help in modeling the number of dengue cases, such as: Poisson multivariate regression prediction models (Hii *et al*., 2012; Sang *et al*., 2014), binomial multivariate linear regression analysis negative (Minh Na; Rocklöv, 2014), Poisson model and its generalizations (Freitas *et al*., 2020) and negative binomial regression approach and Poisson (Chandrakantha, 2019). However, not only the modeling of this phenomenon is important, but also the analysis of the residuals and diagnosis of the model. Thus, the residual examination is an approach that identifies possible general discrepancies between models and observations, and observations that are not accommodated by the models, that is, nonlinear effects, overdispersion, zero inflation and outliers (Scudilio; Pereira, 2019; Feng, Sadeghpour, 2020). Thus, this study aims to analyze the adequacy of Poisson models and their generalizations in relation to the weekly behavior of dengue cases in Recife, from 2009 to 2018, based on the analysis of residuals and diagnosis of the model.

MATERIALS AND METHODS

Data: The observations used for the development of the work refer to the weekly numbers of dengue cases and meteorological variables, to mention: Air Temperature, Air Humidity, Dew Point Temperature, Atmospheric Pressure, Wind Speed and Precipitation in Recife/PE, provided by the Citizen Information Service (CIS) and the National Institute of Meteorology, in the period January 2009 to December 2018. Meteorological variables are collected daily and transformed into weekly scales.

Influence Measures: In Regression analysis there is a step known as residual diagnostic analysis. This analysis is important as it helps detect outliers that deserve further analysis. The verification of possible departures from the assumptions made for the model is a step that deserves a lot of attention when a model is adjusted to a data set, taking into account the random and systematic part of the model, as well as verifying the presence of observations with some influence disproportionate in the results of the adjustment. Influence measures are used to analyze how much a specific observation affects some property of the studied model (Paula, 2004; Cordeiro; Demétrio, 2008). After the elaboration and adjustment of the models, it is necessary to observe the existence of failures in the adjustment of the proposed models, based on the residual analysis and diagnosis of the generalized linear model. Some measures were used to evaluate, namely: Cook's distance, leverage point, Student Residues and half-normal, capable of evaluating whether an observation has an influence on parameter estimates and on the adjustment statistics (Hinde; Demetrius, 1998; Tedeschi, 2006).

Leverage: For the detection, evaluation and understanding of influencing points, the study of the H matrix is used, which is known as the coefficient or regressor matrix or even the hat matrix (Majumbar; Pal, 2016). A fundamental characteristic of the matrix H is intrinsic to the elements $h_{ii}, i = 1, \dots, n$, of its diagonal. The element h_{ii} measures how far an observation, y_i, it's from the other $n-1$ observations in the space defined by the explanatory variables of the model (Cordeiro; Neto, 2004). For linear models the matrix is given by (Dorier, *et al.*, 2020):

$$H = X(X^T X)^{-1} X$$

In the context of generalized linear models (GLMs), observations can be identified by the element h_{ii} of the generalized hat matrix, defined by:

$$\hat{H} = W^{1/2} X (X^T \hat{W} X)^{-1} X^T W^{1/2} \quad (2)$$

in which W is the value of W in $\hat{\beta}$ (Cordeiro; Neto, 2004). If the input values of a hat matrix diagonal are large, the measurement has greater leverage or influence on the estimated states than others and can be called leverage points (Majumbar; Pal, 2016).

Pearson Residuals

Among the most common types of residues observed in GLMs, there is the Pearson residue, which is also the simplest, being defined by:

$$r_i^p = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{V}_i(\hat{\mu}_i)}} \quad (3)$$

in which $\hat{\mu}_i$ and $\hat{V}_i$ are the adjusted mean and adjusted variance function, respectively, of Y_i (Clarice; Demétrio, 2008).

Cook's Distance: Cook's distance is a measure of influence for each observation on parameter estimates. It measures the distance between the estimate of the parametric vector using all observations ($\hat{\beta}$), but eliminating the i-th observation ($\hat{\beta}_i$) (Cook, 2000). Given a matrix X, with n x p known values (explanatory variables), β a vector n x p, unknown parameters, the Cook distance is defined as:

$$D_i = \frac{(\hat{\beta}|(-i) - \beta)' X' X (\hat{\beta}_{(-i)} - \beta)}{ps^2} \quad i = 1, 2, \dots, n \quad (5)$$

in which $\hat{\beta}_{(-i)}$ denotes the least squares estimate of $\hat{\beta}$ with the *i-th* deleted point, p the number of parameters and $s^2 = R'R/(n-p)$, where $R = Y - \hat{Y}$, is the residual vector.

In the context of GLMs, Cook's distance is given by:

$$D_i = \frac{r_i^2}{p} \frac{V(\hat{Y}|i)}{V(R_i)} \quad (6)$$

in which r_i is the studentized residual, p the number of parameters, $V(\hat{Y}|i) = x_i'(X'X)^{-1}x_i\sigma^2 = v_i\sigma^2$, variation of the *i-th* predicted value and $V(R_i) = \sigma^2(1 - v_i)$ variance of the *i-th* residual (Cook, 2000; Jiménez, 2001, Justo; Pintarelli, 2019). So, the higher the value of D_i the greater will be the influence of observation i in the adjusted model.

Half-normal plot: After estimating, all information available for model verification is contained in the residuals. The distribution assumption for a generalized linear model is often verified by plotting the ordered absolute values of a diagnostic measure *versus* the corresponding expected values of the order statistics, in absolute value, of the half-normal distribution. The use of the half-normal plot with simulated envelope to verify the goodness of fit of the model, especially when the data present overdispersion (Mccullagh; Nelder, 2019; Hinde; Demetrius, 1998)

RESULTS AND DISCUSSION

Based on the significant parameters of the proposed models (Table 1), it was detected that the Poisson model presented overdispersion ($\phi >$ 1.00). According to Zeviani *et al*. (2014), overdispersion will affect parameter estimates, making them inefficient, with inconsistent errors and higher error rate for hypothesis testing applied to non-equidisperse data. Thus, to satisfactorily accommodate the overdispersion, other models were adjusted, such as Quasi-Poisson and Negative Binomial.

Table 1. Significant parameters of the Poisson, Quasi-Poisson and Negative Binomial regression models for the number of dengue cases grouped weekly in Recife, Brazil, from 2009 to 2018

Time		*Models*					
		Poisson	*VIF*	*Quasi-Poisson*	*VIF*	*Binomial Negativa*	*VIF*
Period							
2009 a 2018	α	-73.356	-	12.922	-	14.391	-
	β_1	1.001	1.26	-0.081	1.13	0.103	1.14
	β_2	0.154	1.13	-	-	-	-
	β_3	-0.434	1.26	-	-	-	-
	β_4	0.053	1.17	-	-	-	-
	β_5	-1.006	1.08	-	-	-	-
	β_6	-8.10^{-5}	1.15	5.10^{-4}	1.13	7.10^{-4}	1.14
	ϕ	139.8		212.1		1.79	

α= intercept; β_1= Air humidity; β_2= Temperature; β_3= Dew point temperature; β_4= Atmospheric pressure; β_5= Wind speed; β_6= Precipitation; ϕ =Dispersion parameter; VIF = Variance inflation factor.

Among the three models studied, the Negative Binomial model showed the best fit to the expected number of dengue cases in Recife, in which the dispersion parameter (ϕ= 1.79) was lower compared to the Poisson and Quasi-Poisson models in the period from 2009 to 2018, corroborating the results found by Chandrakantha (2019), where the Negative Binomial model showed better accuracy due to the nature of the overdispersed data. The variables relative humidity (β_1) and precipitation (β_6) are the most frequent and significant predictors. Previously, Freitas *et al*. (2020) also demonstrated a direct relationship of dengue cases with meteorological factors, where they observed positive/negative estimates of temperature and adverse effects of precipitation with dengue cases in Recife. The relative humidity of the air was one of the significant factors that most affected the incidence of dengue. For every 1 unit increment in relative air humidity, the log of expected change in the number of cases increased by 10,84% ($e^{0,103}$ = 1.108), maintaining the other variables constant. Extensive study by Naish *et al*. (2014), demonstrates that relative humidity and air temperature are the two most important predictors, with a potential impact on the life cycle of mosquitoes at different stages. It was observed that all model parameters have a VIF less than 5 indicating a low correlation between the predictors, that is, a measure that disturbs the GLM estimates (Walker; Birch, 1988). In Table 2, it can be seen that the Negative Binomial regression model has lower AIC, BIC and NDE compared to the other models, therefore, it is much more parsimonious and performed better to estimate the number of dengue cases in the region. However, it appears that the coefficient of determination of the Negative Binomial model is always lower than the Poisson and Quasi-Poisson model.

Table 2: Information criteria, NDE and coefficient of determination of Poisson, Quasi-Poisson and Negative Binomial regression models for the number of dengue cases grouped weekly in Recife, Brazil, from 2009 to 2018

	Recife Models: 2009 - 2018		
	Poisson	Quasi-Poisson	Negative Binomial
AIC	8840.4	-	834.9
BIC	8854.1	-	841.9
NDE	160.9	226.27	1.06
	0.68	0.55	0.50

By means of the half-normal graph, in fact, it is verified that practically all points are outside the simulated envelope, with a confidence level of 95%, showing that the Poisson model presented an unsatisfactory fit (Figure 1A). A result also observed in the Quasi-Poisson model (Figure 1B), since part of the residues are outside the simulated envelope. On the other hand, it was found that the Negative Binomial model (Figure 1C) had better performance; all points are within the simulated envelope except for one point at the end of the graph. And from the Shapiro-Wilk test, the null hypothesis was not rejected (Shapiro; Wilk, 1965), that is, the normality of the errors was achieved in the Quasi-Poisson and Negative Binomial models (p-value = 0.1387 and 0.094, respectively) adopting a significance level of 5%, allowing for more precise inferences and practical interpretations for the parameters of the model.
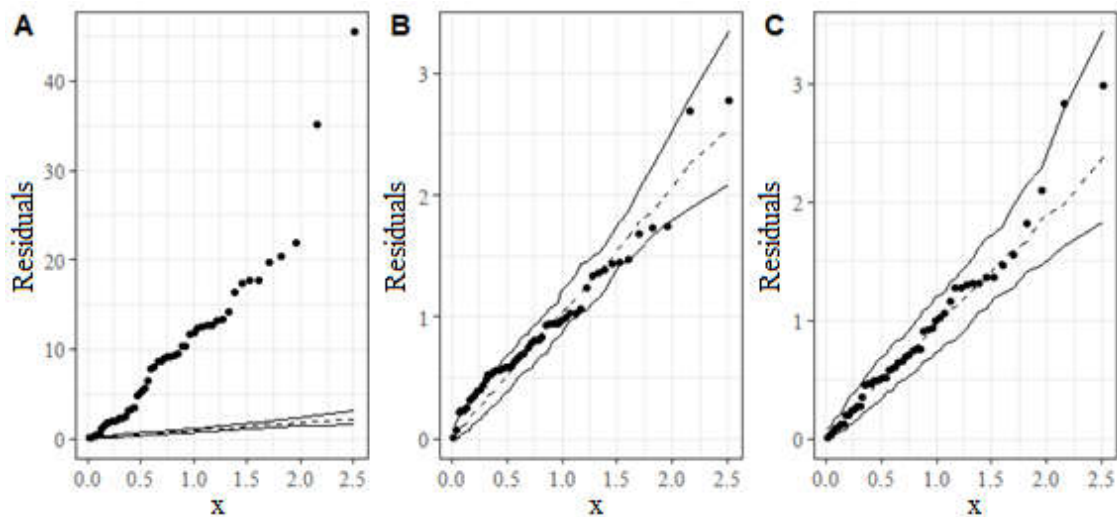
Figure 1: Half-normal graph for the Poisson (A), Quasi-Poisson (B) and Negative Binomial (C) model for the number of dengue cases grouped weekly in Recife, Brazil, in the period from 2009 to 2018

According to Bruce and Bruce (2017), a data point has high leverage if the value of this statistic is above $2(p+1)/n$, in which p is the number of predictors and n is the number of observations. In Figure 2, graphs for projection H (Figure 2A), Cook's distance (Figure 2B) and Pearson's residual (Figure 2C) are shown for the Negative Binomial model. It can be seen in Figure 1A that there are three high leverage points in the data (EW 24, 26 and 32) which may have influenced against a regression line. Venables and Ripley (2013) demonstrate that transformation in the response variable can circumvent the effect of outliers in the model. The graph of Cook's Distance (Figure 2B) *versus* the order of observations (Epidemiological week) revealed the presence of two observations (EW 24 and 51) considered influential. The re-accommodation of the model to the data without the presence of these observations did not alter the inferential results. According to Paula (2004), the reason for finding atypical values in relation to the others does not necessarily imply in the elimination of such points, this being the last resource to be used.
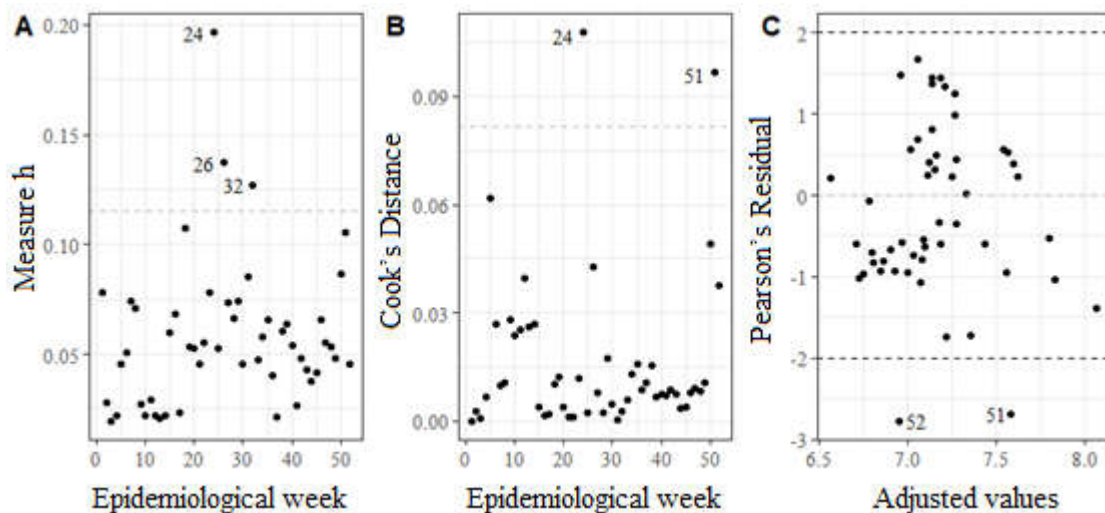

Figure 2. Results of diagonal elements of the projection matrix H (A), Cook's distance (B) and Pearson's residual (C) for the Negative Binomial model, for the number of dengue cases grouped weekly in Recife, Brazil, in 2009 to 2018.

In Figure 2C, it can be seen that the residual graph does not present any adjusted pattern, in which the presence of a pattern may indicate a problem with some aspect of the linear model. Therefore, it was possible to assume a linear relationship between predictors and the response variable. Furthermore, as shown in Figure 2C, it is noted that the residues are satisfactorily distributed between -2 and 2, with the exception of EW 51 and 52. When relating the three graphs, it was observed that EW 51 is an influential point and EW 24 is the atypical data point, with EW 24 located in the period with the highest rainfall in the municipality. According to Lowe *et al.* (2011), from the GLM there was an indication that the climatic covariates play a statistically significant role in the transmission of dengue.

CONCLUSION

Taking into account the criteria adopted in this work to verify the goodness of fit of the models, the Negative Binomial model was the most suitable for the data. When performing residual analysis and diagnostic techniques, it was observed the existence of atypical and influential values, in which the behavior of such events were impacted by other predictors and the dependence was captured using a regression structure. The diagnostic tools were able to identify a multicollinearity problem, requiring the removal of regressors in the modeling. Therefore, the diagnosis is essential for validating a model that aims to interpret and understand the results.

Finally, for future work, we will use georeference tools in order to map the other municipalities in the state of Pernambuco, applying a temporal delay in the precipitation variable. Thus, studying the relationship between rainfall and number of cases helps to guide public policies aimed at health in the state of Pernambuco.

Acknowledgment

REFERENCES

Augustin, N. H., Sauleau, E. A., & Wood, S. N. (2012). On quantile quantile plots for generalized linear models. *Computational Statistics & Data Analysis*, *56*(8), 2404-2409.https://doi.org/10.1016/j.csda.2012.01.026.

Bhatt S, Gething PW, Brady OJ, Messina JP, Farlow AW, Moyes CL, Drake JM, Brownstein JS, Hoen AG, Sankoh O, Myers MF, George DB, Jaenisch T, Wint GR, Simmons CP, Scott TW, Farrar JJ, & Hay SI (2013). The global distribution and burden of dengue. *Nature*, 496(7446), 504-507. https://doi.org:10.1038/nature12060.

Bruce, P., & Bruce, A. (2017). Practical statistics for data scientists: 50 essential concepts. " O'Reilly Media, Inc.".

Chandrakantha, L. (2019). Statistical analysis of climate factors influencing dengue incidences in Colombo, Sri Lanka: Poisson and negative binomial regression approach. *Int. J. Sci. Res. Publ*, *9*, 133-144.http://dx.doi.org/10.29322/IJSRP.9.02.2019.p8616.

Cook, R. D. (2000) Detection of Influential Observation in Linear Regression, Technometrics, 42:1, 65-68. https://doi.org/10.1080/00401706.2000.10485981.

Cordeiro, G. M., & Neto, E. D. A. (2004). Modelos paramétricos. *Pernambuco: UFRPE*.

Cordeiro, G. M., & Demétrio, C. G. (2008). Modelos lineares generalizados e extensões. *Piracicaba: USP*.

de Freitas, J. R., de Miranda Oliveira, M. G. F., Cunha Filho, M., da Silva, F. S. G., & de Vasconcelos, J. M. (2020). Modelo de Poisson e suas generalizações aplicadas a dados de dengue, Brasil. *Research, Society and Development*, *9*(10), e6629108874-e6629108874.https://doi.org/10.33448/rsd-v9i10.8874.

do Carmo, R. F., Júnior, J. V. J. S., Pastor, A. F., & de Souza, C. D. F. (2020). Spatiotemporal dynamics, risk areas and social determinants of dengue in Northeastern Brazil, 2014–2017: an ecological study. *Infectious Diseases of Poverty*, *9*(1), 1-16.https://doi.org/10.1186/s40249-020-00772-6.

do Nascimento, I. D. S., Pastor, A. F., Lopes, T. R. R., Farias, P. C. S., Gonçales, J. P., do Carmo, R. F., ... & Júnior, J. V. J. S. (2020). Retrospective cross-sectional observational study on the epidemiological profile of dengue cases in Pernambuco state, Brazil, between 2015 and 2017. *BMC Public Health*, *20*(1), 1-10. https://doi.org/10.1186/s12889-020-09047-z.

Dorier, M., Frigo, G., Abur, A., & Paolone, M. (2020). Leverage Point Identification Method for LAV-Based State Estimation. *arXiv preprint* arXiv:2008.03941.

Feng, C., Li, L., & Sadeghpour, A. (2020). A comparison of residual diagnosis tools for diagnosing regression models for count data. *BMC Medical Research Methodology*, *20*(1), 1-21. https://doi.org/10.1186/s12874-020-01055-2.

Ferreira, A. C., Chiaravalloti Neto, F., & Mondini, A. (2018). Dengue em Araraquara, SP: epidemiologia, clima e infestação por Aedes aegypti. *Revista de Saúde Pública*, *52*, 18. https://doi.org/10.11606/S1518-8787.2018052000414.

Hii, Y. L., Zhu, H, Ng. N., Ng, L. C., & Rocklöv, J. (2012) . Forecast of Dengue Incidence Using Temperature and Rainfall. *PLoS Neglected Tropical Diseases*; 6(11):e1908. pmid:23209852.https://doi.org/10.1371/journal.pntd.0001908.

Hinde, J., & Demétrio, C. G. B. (1998). Overdispersion: models and estimation. *Computational statistics & data analysis*, *27*(2), 151-170. https://doi.org/10.1016/S0167-9473(98)00007-3.

Jiménez, J. (2001). Una Generalización de la estadística de Cook. *Revista Colombiana de Estadística*, *24*(2), 111-120.

Justo, C. E., Pintarelli, M. B. (2019). Análisis en los residuos de redes altimétricas mínimamente condicionadas. In *V Jornadas de Investigación, Transferencia y Extensión de la Facultad de Ingeniería (La Plata, 2019)*. https://revistas.udistrital.edu.co/index.php/azimut/article/view/16539.

Lowe, R., Bailey, T. C., Stephenson, D. B., Graham, R. J., Coelho, C. A., Carvalho, M. S., & Barcellos, C. (2011). Spatio-temporal modelling of climate-sensitive disease risk: Towards an early warning system for dengue in Brazil. *Computers & Geosciences*, *37*(3), 371-381.https://doi.org/10.1016/j.cageo.2010.01.008.

Majumdar, A., & Pal, B. C. (2016). Bad data detection in the context of leverage point attacks in modern power networks. *IEEE Transactions on Smart Grid*, *9*(3), 2042-2054. https://doi.org/10.1109/TSG.2016.2605923.

McCullagh, P., & Nelder, J. A. (2019). *Generalized linear models*. Routledge, Cap. 2.

Minh An, D. T., & Rocklöv, J. (2014). Epidemiology of dengue fever in Hanoi from 2002 to 2010 and its meteorological determinants. *Global health action*, *7*(1), 23074. https://doi.org/10.3402/gha.v7.23074.

Naish, S., Dale, P., Mackenzie, J. S., McBride, J., Mengersen, K., & Tong, S. (2014). Climate change and dengue: a critical and systematic review of quantitative modelling approaches. *BMC infectious diseases*, *14*(1), 1-14.http://www.biomedcentral.com/1471-2334/14/167.

Nunes, P. C. G., Daumas, R. P., Sánchez-Arcila, J. C., Nogueira, R. M. R., Horta, M. A. P., & dos Santos, F. B. (2019). 30 years of fatal dengue cases in Brazil: a review. BMC Public Health, 19:329.https://doi.org/10.1186/s12889-019-6641-4.

Oliveira, R. D. M. A. B., Araújo, F. M. D. C., & Cavalcanti, L. P. D. G. (2018). Aspectos entomológicos e epidemiológicos das epidemias de dengue em Fortaleza, Ceará, 2001-2012. *Epidemiologia e Serviços de Saúde*, *27*, e201704414. https://doi.org/10.5123/S1679-49742018000100014.

Paula, G. A. (2004). *Modelos de regressão: com apoio computacional* (pp. 28-55). São Paulo: IME-USP.

Sang, S., Gu, S., Bi, P., Yang, W., Yang, Z., Xu, L., ... & Liu, Q. (2015). Predicting unprecedented dengue outbreak using imported cases and climatic factors in Guangzhou, 2014. *PLoS neglected tropical diseases*, *9*(5), e0003808.https://doi.org/10.1371/journal.pntd.0003808.

Scudilio, J., & Pereira, G. H. A. (2019). *Adjusted quantile residual for generalized linear models. Computational Statistics*. https://doi.org/10.1007/s00180-019-00896-w.

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). Biometrika, 52(3/4), 591-611.

Sobral, M. F. F., & Sobral, A. I. G. D. P. (2019). Casos de dengue e coleta de lixo urbano: um estudo na Cidade do Recife, Brasil. *Ciência & Saúde Coletiva*, *24*, 1075-1082. https://doi.org/10.1590/1413-81232018243.10702017.

Tedeschi, L. O. (2006). Avaliação da adequação dos modelos matemáticos. *Agricultural systems* , *89* (2-3), 225-247. https://doi.org/ 10.1016/j.agsy.2005.11.004.

Walker, E., & Birch, J. B. (1988). Influence measures in ridge regression. *Technometrics*, *30*(2), 221-227. https://doi.org/10.1080/ 00401706.1988.10488370.

Venables, W. N., & Ripley, B. D. (2013). *Modern applied statistics with S-PLUS*. Springer Science & Business Media.

World Health Organization (WHO). *Global strategy for dengue prevention and control 2012-2020* Genebra: WHO; 2012.

World Health Organization (WHO). (2020). Dengue and severe dengue. Geneva.

Zeviani, W. M., Ribeiro Jr, P. J., Bonat, W. H., Shimakura, S. E., & Muniz, J. A. (2014). The Gamma-count distribution in the analysis of experimental underdispersed data. *Journal of Applied Statistics*, *41*(12), 2616-2626. https://doi.org/10.1080/02664763.2014.922168

****