Research, Society and Development, v. 9, n. 11, e71291110299, 2020
(CC BY 4.0) | ISSN 2525-3409 | DOI: http://dx.doi.org/10.33448/rsd-v9i11.10299

Multivariate regression analysis in the probability of deaths in COVID-19 cases: a case study in the State of Pará, Amazon region, Brazil

Análise de regressão multivariada na probabilidade de óbitos em casos COVID-19: um estudo de caso no Estado do Pará, região amazônica, Brasil

Análisis de regresión multivariante en la probabilidad de muerte en casos de COVID-19: un estudio de caso en el Estado de Pará, región amazónica, Brasil

Received: 11/17/2020 | Reviewed: 11/24/2020 | Accept: 11/26/2020 | Published: 12/02/2020

Cássio Pinho dos Reis
ORCID: https://orcid.org/0000-0002-2211-2295
Federal University of Mato Grosso of South, Brazil
E-mail: cassio.reis@ufms.br
Herson Oliveira da Rocha
ORCID: https://orcid.org/0000-0002-2494-6277
Federal Rural University, Brazil
E-mail: herson@ufra.edu.br
Nayara de Araújo Muzili Reis
ORCID: https://orcid.org/0000-0001-6809-7951
Municipal Health Secretariat, Brazil
E-mail: naymuzili@gmail.com
Sávio Pinho dos Reis
ORCID: https://orcid.org/0000-0001-6919-3638
State University of Pará, Brazil
E-mail: savio.reis@uepa.br
Gustavo Nogueira Dias
ORCID: https://orcid.org/0000-0003-1315-9443
Colégio Federal Ten. Rêgo Barros, Brazil
E-mail: gustavonogueiradias@gmail.com
Gilberto Emanoel Reis Vogado
ORCID: https://orcid.org/0000-0003-4763-4767
Universidade do Estado do Pará, Brazil
E-mail: gvogado@globo.com

Vanessa Mayara Souza Pamplona
ORCID: https://orcid.org/0000-0002-2461-2103
Universidade Federal Rural da Amazônia, Brazil
E-mail: vanessamayara2@gmail.com
Washington Luiz Pedrosa da Silva Junior
ORCID: https://orcid.org/0000-0002-1413-0047
Colégio Federal Ten. Rêgo Barros, Brazil
E-mail: jwl_pedrosa@hotmail.com

Abstract

Since the first detected cases of COVID-19 in Brazil, researchers have made a great effort to try to understand the disease. Understanding the impact of the disease on people can be instrumental in identifying which groups can be considered at risk. Therefore, this study researches a probabilistic model based on a statistical model of non-linear regression analyzing the following variables: age, if you are a health professional, if you are resident in the Metropolitan Region of Belém (RMB), State of Pará and gender with the objective of identifying those people who have a greater impact on the number of people infected and killed by COVID-19, that is, people who are more likely to die. To carry out the research, we used the data of all infected people by COVID-19 in the State of Pará until July 2020. It can be verified according to the proposal of the probabilistic model that elderly people, with a odds ratio of 1.69 (95% CI 1.52-1.88), residents of Metropolitan Region of Belém, with an odds ratio of 2.14 (95% CI 2.02 - 2.27) and men, with an odds ratio of 1.83 (95% CI 1.73 - 1.95) are groups of people with a higher risk of dying from diseases, while health professionals, with a 0.36 chance ratio (CI9 5% 0.29 - 0.45), are less likely to die.

Keywords: Probabilistic model; COVID-19 in Brazil; Risk group; Amazon region.

Resumo

Desde os primeiros casos detectados de COVID-19 no Brasil, os pesquisadores têm feito um grande esforço para tentar entender a doença. Compreender o impacto da doença nas pessoas pode ser fundamental para identificar quais grupos podem ser considerados de risco. Diante disso, este estudo pesquisa um modelo probabilístico baseado em um modelo estatístico de regressão não linear analisando as seguintes variáveis: idade, se você é profissional de saúde, se é residente na Região Metropolitana de Belém (RMB), Estado do Pará e gênero com o objetivo de identificar aquelas pessoas que têm um maior impacto no número de infectados e de óbitos por COVID-19, ou seja, pessoas com maiores probabilidades de ir a óbito. Para a realização da pesquisa, utilizamos os dados de todas as pessoas contaminadas pelo COVID-19 no Estado do Pará até julho de 2020. Pode ser verificado de acordo com a proposta do modelo probabilístico que idosos, com razão de chance de 1,69 (IC95% 1,52-1,88), moradores da Região Metropolitana de Belém, com razão de chance de 2,14 (IC95% 2,02 – 2,27) e os homens, com razão de chance de 1,83 (IC95% 1,73 – 1,95) são grupos de pessoas com maior risco de morrer de doenças, enquanto que profissionais da saúde, com razão de chance de 0,36 (IC95% 0,29 – 0,45), apresentam menores probabilidades de ir a óbito.

Palavras-chave: Modelo probabilístico; COVID-19 no Brasil; Grupo de risco; Região Amazônica.

Resumen

Desde los primeros casos detectados de COVID-19 en Brasil, los investigadores han hecho un gran esfuerzo para intentar comprender la enfermedad. Comprender el impacto de la enfermedad en las personas puede ser fundamental para identificar qué grupos pueden considerarse en riesgo. Por tanto, este estudio investiga un modelo probabilístico basado en un modelo estadístico de regresión no lineal analizando las siguientes variables: edad, si es un profesional de la salud, si es residente en la Región Metropolitana de Belém (RMB), Estado de Pará y sexo con el objetivo de identificar a aquellas personas que tienen un mayor impacto en la cantidad de personas infectadas y asesinadas por COVID-19, es decir, las personas que tienen más probabilidades de morir. Para realizar la investigación, utilizamos los datos de todas las personas infectadas por COVID-19 en el Estado de Pará hasta julio de 2020. Se puede verificar según la propuesta del modelo probabilístico que las personas mayores, con una razón de momios de 1,69 (IC 95% 1,52-1,88), residentes de La Región Metropolitana de Belém, con un odds ratio de 2,14 (IC 95% 2,02 - 2,27) y los hombres, con un odds ratio de 1,83 (IC 95% 1,73 - 1,95) son grupos de personas con mayor riesgo de morir por enfermedades, mientras que los profesionales de la salud, con una razón de probabilidad de 0,36 (IC9 5% 0,29 - 0,45), tienen menos probabilidades de morir.

Palabras clave: Modelo probabilístico; COVID-19 en Brasil; Grupo de riesgo; Región Amazónica.

1. Introduction

The World Health Organization (WHO) declared the COVID-19 epidemic constituted a Public Health Emergency of International Importance (ESPII) is a pandemic, (Oliveira et al., 2020). Severe acute respiratory syndrome coronavirus (SARS-CoV-2) is the etiologic agent of coronavirus 19-induced disease (COVID-19) that emerged in China in late 2019, (Zhou et al.,2020).

The main symptoms caused by SARS-CoV-2 infections are respiratory symptoms, fever, cough, shortness of breath and breathing difficulties, although some infected people have been declared asymptomatic (Petropoulos, et al., 2020). And in the most severe cases, the infection can cause severe acute respiratory syndrome, pneumonia, kidney failure, hospitalization and death.

From the detection of rumors about the emerging disease in Brazil, the Emergency Operations Center (COE) was activated. This organ is linked to the Ministry of Health (MS), coordinated by the Health Surveillance Secretariat (SVS/MS) and, to plan, organize and implement actions with the actors involved in order to monitor the epidemiological situation, developing a contingency plan and control actions such as lockdowns (Oliveira et al., 2020).

In order to date, there is still no specific treatment for COVID-19 disease caused by the new coronavirus. Nevertheless, several researchers from different parts of the world are committed to developing a vaccine, starting to conduct clinical trials in humans such as China with the “Ad5-nCoV” vaccine, the United Kingdom with “ChAdOx1 nCoV-19”, Cuba with "Sovereign 01" and Russia with "Sputnik V", among other countries.

However, symptoms can be treated, and some supportive care for infected people can be highly effective. The following prevention practices are recommended: avoid touching the face, washing hands with water, soap or using gel alcohol, not sharing personal objects, covering the nose and mouth when sneezing or coughing, keeping the rooms well ventilated and ventilated, avoid crowding, especially if you have symptoms or sick, in case of fever, cough and difficulty breathing, seek medical attention quickly, and social isolation to avoid high spread (Cabral et al., 2020) .

In Brazil, since the first cases detected, several authors carried out a task force to try to understand the virus in the most diverse areas such as biomedicine, infectious diseases, mathematics, biophysics, logistics, health, among others. In early March 2020, (Jesus, J.G. et al., 2020), presented a brief report and phylogenetic analysis of confirmed COVID-2 cases in Brazil in two patients from Italy. Detailed clinical and epidemiological descriptions for suspected and confirmed patients are available on the National Public Health Emergency Response and Response Network of the brasilian Ministry of Health of Brazil (Croda, et al., 2020).

The confirmatory real-time RT-PCR tests were performed at the Institute Adolfo Lutz (IAL), a regional reference laboratory for virus detection in the state of São Paulo, Brazil. The SPBR1 patient sample had an RT-PCR cycle threshold value (Ct) of, while infection of the the SPBR2 patient by virus was also confirmed by RT-PCR in real time, sending the viral RNA to IAL for genomic sequencing. The SPBR2 patient sample had an RT-PCR Ct value of, (Jesus, J.G. et al., 2020).

The authors (Candido, et. al., 2020), used travel data on all air travels that had a Brazilian city as their final destination during February and March 2019 as a proxy for flight density during the COVID-2019 outbreak. They focused on data from countries that had reported cases of SARS-CoV-2 by March 5, 2020. The survey collected the total number of passengers flying to Brazilian airports during this period, size of the country's population by 2019 (from the World Population Prospects 2019 database) and the number of COVID-19 cases reported by WHO. The results showed the continuous integration of these data flows should help to guide the resources implementation to mitigate the COVID-19 transmission.

According to (Werneck & Carvalho, 2020) the response to the COVID-19 pandemic can be subdivided, schematically and simplistically, into four phases. They are: containment, mitigation, suppression and recovery. The first phase involves, mainly, the active tracking of passengers coming from abroad and their contractors, describes the avoidance or postponement of community transmission.

The mitigation phase begins when the sustained transmission of the infection is already installed in the country. In this case, the objective here is to decrease the levels of disease transmission for the groups most at risk of having severe clinical conditions, in addition, of course, to the isolation of the positive cases identified. These measures, called "vertical isolation", are generally accompanied by some degree of reduction in social contact.

The suppression phase may be necessary when the previous measures fail to be effective, either because their implementation cannot be carried out properly and immediately or because the achieved reduction in transmission is insufficient to prevent the collapse in health care. In this phase, it is recommended to implement more radical measures of social distance for the entire population, such as, for example, the lockdown.

The recovery phase is when there is a consistent sign of involution of the epidemic and the cases number becomes residual. This last phase requires an organization of society for the social and economic restructuring of the country.

The initial analyzes of the new coronavirus evolution need to be of a regional order, in order to reduce errors and avoid nonsense in analysis and inference (Cabral, et al., 2020). Thereby, they proposed to analyze the evolution data of the new coronavirus in the state of Pará, in addition to identifying a more adequate 75 inference model to estimate the number of infected people in the state.

In the state of Amapá, the authors (Dias, N. L et al., 2020), presented an analysis of SARS-CoV-2 spread using three approaches. Initially, they applied the Imperial College London (ICL) model for the pandemic in Brazil, aiming the implementation of a linear comparative projection for the Amapá population. The second approach was developed with the standard short-term solution of the Susceptible-Infected-Recovered (SIR) model, where it was shown the typical exponential behavior satisfactorily describes the data for the first weeks of the epidemic. However, soon after, there are early discrepancies as a result of a sudden deceleration in the temporal evolution of the number of cases due to isolation measures.

This new regime is appropriately described with the third approach based on the vSIR model, which is a variant of the SIR model. The results presented enable, on the one hand, a better understanding of the scenarios already faced by the population and on the other hand provide short-term projections that will be constantly updated (Dias, N. L et al., 2020).

The epidemiological SIR model proposed by (Kermack, W. O. et al., 1927), was adapted by (Cooper, et. al., 2020) to model approach on the pandemic due to the spreading of the new COVID-19 to investigate the evolution tie of different populations and to monitor diverse significant parameters for the disease spread in various communities, represented by China, South Korea, India, Australia, USA and Italy

A new SEIR (Susceptible-Exposed-Infected-Recovered) agent-based model that aims to simulate the pandemic dynamics using a society of agents emulating people, business and government, proposes the COVID-ABS (Silva, P. C., et al., 2020).

Seven different scenarios of social distancing interventions were analyzed, with varying epidemiological and economic effects: (1) do nothing, (2) lockdown, (3) conditional lockdown, (4) vertical isolation, (5) partial isolation, (6) use of face masks, and (7) use of face masks together with 50% of adhesion to social isolation. With the impossibility of introducing lockout scenarios that have the lowest number of deaths and the highest effect on the economy, scenarios that combine the use of face masks and partial isolation could be more feasible for social cooperation implementation.

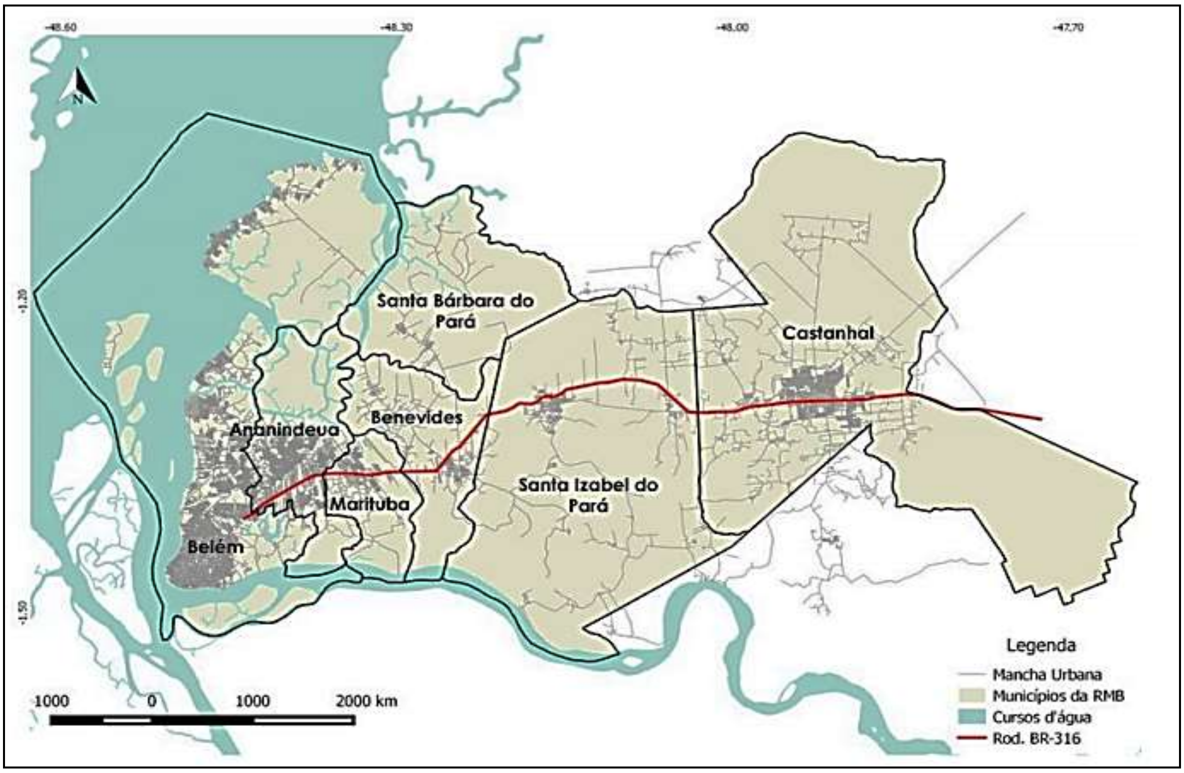
This research proposes a new probabilistic model based on a statistical model of non-linear regression in the Belém metropolitan region, analyzing the following variables: (i) age, (ii) if you are a health professional, (iii) if you are a resident of the Belém metropolitan region (iv) sex, with the aim of identifying those that have the greatest impact on the number of people infected and killed by COVID-19, that is, people who are more probability to die.

2. Study Area

The Belém Metropolitan Region (BMR) was institutionalized at the state level on October 19, 1995, by virtue of Complementary Law 027, formed by the municipalities of Belém, Ananindeua, Marituba, 106 Benevides and Santa Bárbara. By Complementary Law, the municipality of Santa Izabel do Pará was incorporated into the BMR and, by Complementary Law , also the municipality of Castanhal (Fig. 1) (Oliveira Ribeiro, 2016).

It can be seen today the BMR is composed of seven cities, articulated in different dynamics of incorporation into urbanization of a metropolitan character. A process of space production that is no longer accompanied by population concentration, but by dispersion, composing a concentrated and dispersed urban region (Mendes, 2018).

Figure 1. Current configuration of the Belém Metropolitan Region (BMR), Brazil.



Source: (Cardoso & Miranda, 2018)

3. Material and Methods

The data used in this research were obtained from the State Department of Health of Pará (SESPA-PA), Brazil, through the COVID-19 portal available on the internet for any user. These data refer to all notified cases of COVID-19 in the State of Pará, in the period between March 9, 2020, when the first cases were recorded, until July 29, 2020. The statistical analyzes were all made using software R. (Chambers, J., 2008).

The regression methods have become integral components of any data analysis, whose interest is to describe the relationship between a response variable and a set of explanatory variables, where the main objective is to find a functional, adequate and parsimonious way to describe the relationship between a response variable (dependent) and a set of independent variables (explanatory)[17]. The most common example of modeling is the simple linear regression model where the response variable is assumed to be continuous, (Pregibon, D. 1981; Allison, P. D., 2012; Menard, S., 2002 & Hosmer J., et al., 2013).

The big difference between a simple linear regression model and the logistic model is the fact that the response variable is binary or dichotomous, and the difference in the adjusted functional form. The logistic model estimates the probability of a certain situation to occur or not, based on certain characteristics. Thus, in this research, we consider the simple linear regression model proposed by (Neter, J. et al., 1983), given by Eq. 1 below:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \tag{1}$$

Where the response variable is binary, that is, it assumes the values 0 or 1 in the absence or presence of the characteristic under study, respectively. The expected answer in Eq. 1 is given by:

$$E(Y_i) = \beta_0 + \beta_1 X_i \tag{2}$$

Thus, consider Y_i a Bernoulli aleatory variable with a probability distribution given by:

$$\begin{cases} Y_i = 1,\ if\ P(Y_i = 1) = \pi_i \\ Y_i = 0,\ if\ P(Y_i = 0) = 1 - \pi_i \end{cases}$$

By definition, we have to $E(Y_i) = \pi i$. Therefore, we have to:

$$E(Y_i) = \beta_0 + \beta_1 X_i = \pi_i \tag{3}$$

Thus, the average response $E(Y_i)$, when the response variable Y_i is a binary variable, always represents the probability of $Y_i = 1$, for the level of the predictor variable X_i. And considering only one independent variable Xi, we have a simple logistic regression model in its usual form is given by Eq. 4:

$$E(Y_i/X_i) = \pi_i = \frac{exp(\beta_0 + \beta_1 X_i)}{1 + exp(\beta_0 + \beta_1 X_i)} \tag{4}$$

where β_0 and β_1 are the regression coefficients to be estimated and Xi is the independent variable, where i = 1, ..., n. In the case of multiple logistic regression, which is an extension of the simple logistic model, because the only difference is that instead of using only one independent variable X_i, now two or more independent variables will be used X_1, X_2,

..., X_n, will now be used, with the model is composed not only of independent variables, but also of regression coefficients β_1, β_2, ..., β_p. So you have, $\beta_{iX} = \beta_0 + \beta_{1Xi,1} + \beta_{2Xi,2} + ... + \beta_{pXi,p}$, where i = 1, ..., n. Therefore, the model described by Eq. 4 extends to the multiple logistic model, given by Eq. 5, below:

$$E(Y_i/X_i) = \pi_i(X_i) = \frac{exp(\beta_i X_i)}{1 + exp(\beta_i X_i)} \tag{5}$$

Therefore, the dependent variable Y_i is given by $Y_i = E(Y_i/X_i) + e_i$, where the term e_i is the model's aleatory error and represents the difference between the observed value of and the conditioned expected value of Y_i given X_i. And the values of parameters β_1, β_2, ..., βp are estiated using the maximum likelihood estimation (MLE). The odds ratio (OR) is one of the main statistics used in the analysis of binary data. It is defined as the ratio between the chance of an event occurring in one group and the chance of occurring in another group (Bland & Altman, 2000). Chance being the probability of an event occurring divided by the probability of the event not occurring, (Agresti, A. 1980 & Szumilas, M. 2010).

The chance is defined as:

$$\frac{P(Y_i = 1/X_i)}{P(Y_i = 0/X_i)} = \frac{\pi(X_i)}{1 - \pi(X_i)} \tag{6}$$

Therefore, the OR is obtained by Eq. 7:

$$OR = \left[\frac{\pi X_i = 1)}{1 - \pi X_i = 1)}\right] / \left[\frac{\pi(X_i = 0)}{1 - \pi(X_i = 0)}\right] \tag{7}$$

To identify the variables that do not have a good fit in estimating the model parameters, there are some tests to select these variables, such as the Stepwise test, (Sarkar, S. K., 2002). This test allows selecting variables from an initial set of explanatory variables. The choice of variables is based on a heuristic procedure, but does not guarantee, from a practical point of view, that the model is the best. However, stepwise is useful in the early stages of analysis, especially when there is a very large number of possible explanatory variables, (Romano, J. P., & Wolf, M., 2005).

In the stepwise it is necessary to establish the probability of entry and exit of the variables in the model and, commonly, 0.20 is used as input probability and 0.05 as output probability. To decide which logistic regression model will be used, it is necessary to apply some validation tests for this model, because without them, the equations may not be statistically reliable. It is also necessary to check if there are influential points (outliers), if the response function is monotonic and in the form of an S (sigmoidal), and if the adjusted logistic model is adequate, (Giancristofaro, R. A., Salmaso, L., 2007 & Lewis-Beck, C.,1984).

In order to validate the model used in this work, the Hosmer-Lemeshow, Pearson and Deviance tests were used. Assessing the quality of fit in logistic regression models can be problematic, since the commonly used deviation does not have approximate distributions, under the null hypothesis of no lack of fit, when continuous covariates are modeled, (Pulkstenis, E., & Robinson, T. J. 2002). The Pearson test measures how much the observation is predicted by the model, the residual of the deviance is a measure of how well the observation is predicted by the model and the Hosmer-Lemeshow test evaluates the adjusted model, comparing the observed and expected frequencies.

The Hosmer-Lemeshow test is a test that evaluates the adjusted model by comparing the observed and expected frequencies. The test associates the data with your estimated probabilities from the lowest to the highest, then takes a chi-square test to determine whether the observed frequencies are close to the expected frequencies, (Xie, X. J. et al., 2002). The Pearson test, on the other hand, is used to analyze the residuals for logistic models. It is a useful measure to assess how well the selected model has adjusted to the data.

The model’s Deviance is a kindness statistic based on the maximized log-likelihood functions to verify whether a subset of the X variables can be removed from the multiple logistic regression model, testing whether the βj regression coefficients are equal to zero, (Qin, X. et al., 2020). Given a sequence of embedded models, Deviance is used as a measure of model discrepancy. It is also very useful to assess how well the selected model fits the data, the higher the descriptive level (p) associated with it, the better the model fits the data.

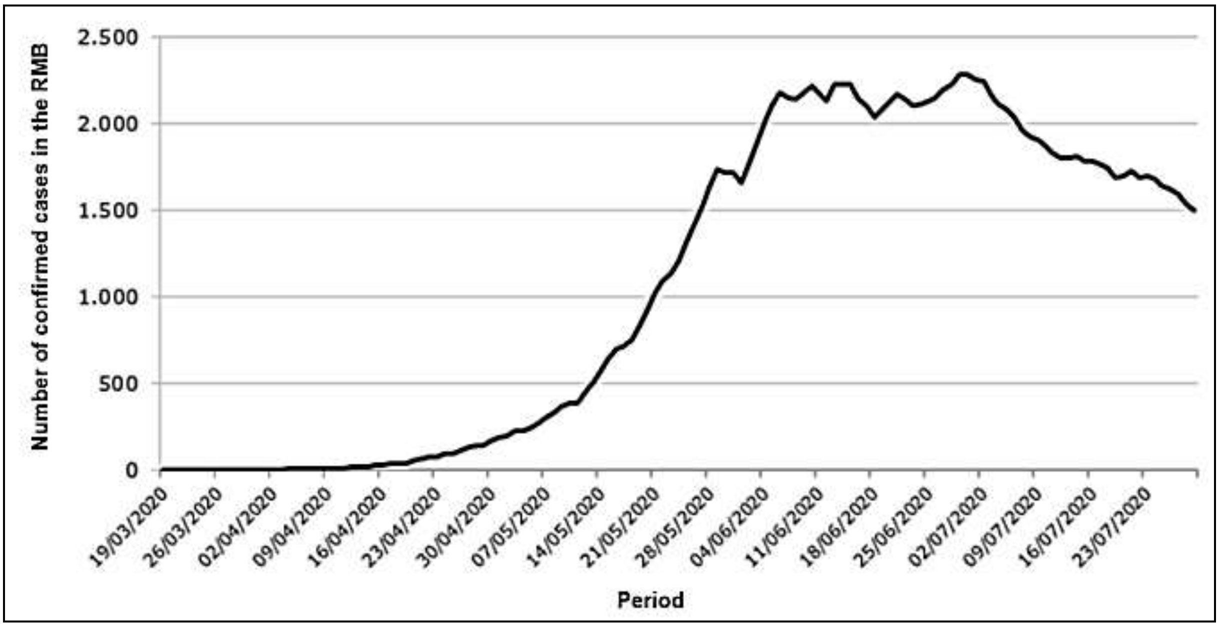
Thus, as a rule of decision of the tests, there is the descriptive level p, which is the probability of occurring values of the test statistic more extreme than that observed, under the null hypothesis (Ho being true. When p is greater than or equal to the level of significance α = 0.05), the null hypothesis is rejected (Priyadarshan, P. M., 2019 & Von Kistowski, J. , 2020).

4. Results

The first confirmed case of COVID-19 in the state of Pará, Brazil was on March 18, 2020. Since then, until July, 151,849 cases have been confirmed. This means that, four months after the beginning of the contamination, 1.72% of the entire population of Pará was diagnosed with the disease. Of this total, 5,694 people died due to COVID-19, that is, a lethality rate of 3.74%. Thus, we can say that for 188 each group of 10,000 people in the state of Pará, 6 died after contracting the disease.

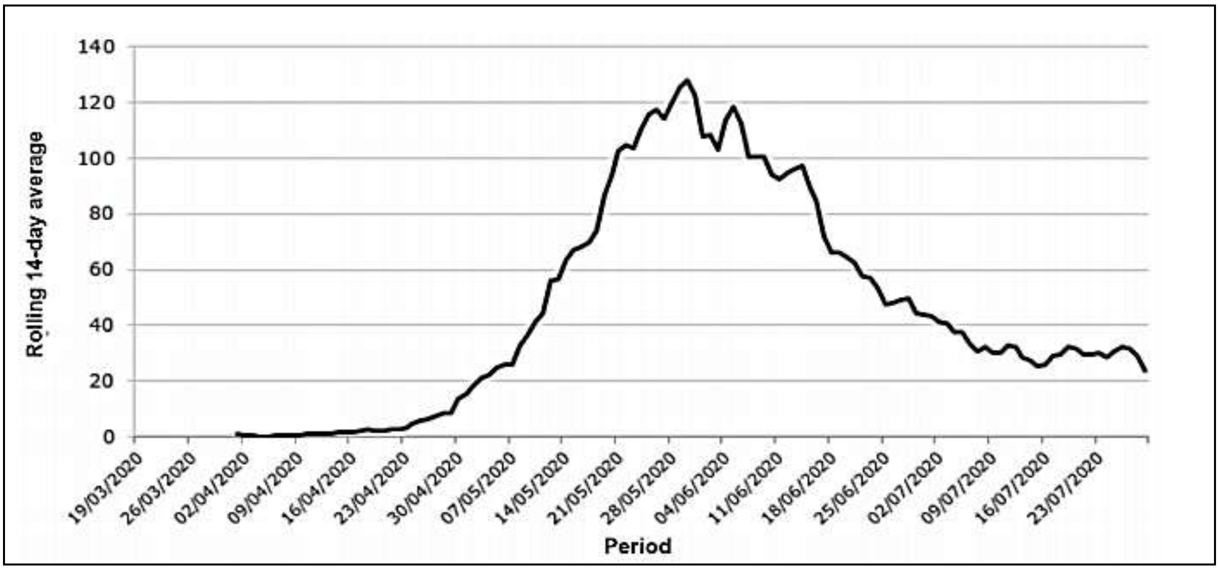
For this reason, one of the ways to observe the disease behavior is by analyzing the number of cases and the moving average of deaths confirmed by COVID-19. In Figs. 2 and 3, it is observed that, between May 21 and June 8, 2020, the moving average of the number of deaths was greater than per day. After this period, it can be seen that the daily average dropped about 80% reaching a moving average of 24 deaths daily. Based on this, it can be inferred that the state of Pará, after a scenario of increase in the number of confirmed cases, begins to enter a stage of stability. The 14-day moving average is the simple arithmetic average of the last 14 days, for example, the information from 15 days ago is not taken into account in today's 14-day moving average.

Figure 2. Number of confirmed cases in metropolitan region of the city of Belém in the period between 03/18/2020 to 07/23/2020.



Source: Authors.

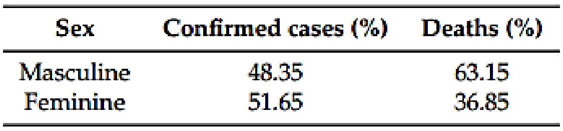
Figure 3. Average of 14 days of confirmed cases in metropolitan region of the city of Belém in the period between 03/18/2020 to 07/23/2020.



Source: Authors.

From the data in Table 1, it appears the majority of confirmed cases of the disease occurred in women (51.65%), however, the majority of deaths occurred in men. This can show men are more resistant when seeking medical help when necessary, causing the disease to worsen and consequently 198 delaying the start of treatment, (Martins, L. K. et al., 2020).

Table 1. Percentage of confirmed cases and deaths by COVID-19 in the state of Pará, Brazil by gender.

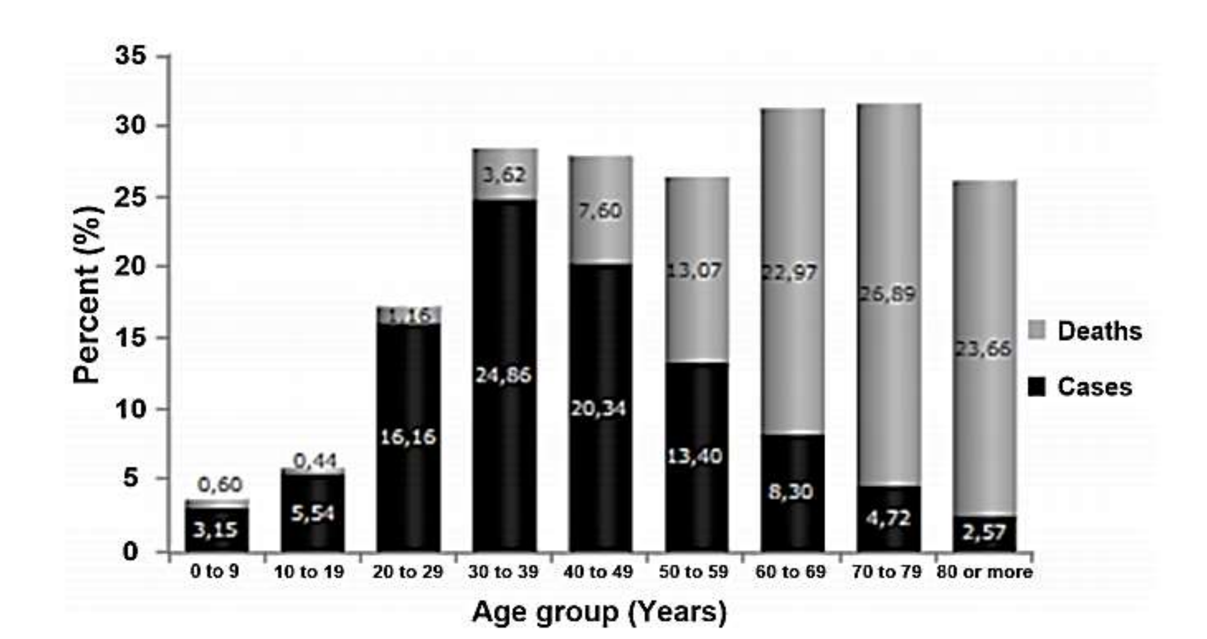


Sex	Confirmed cases (%)	Deaths (%)
Masculine	48.35	63.15
Feminine	51.65	36.85

Source: Authors.

According to the 2010 demographic census, 7.5% of the Pará population are aged between 30 and 39 years, (IBGE, 2010). A large part of the people in this age group are economically active, therefore, they need to travel a lot from their homes to work, increasing the chances of contagion. That may explain the fact almost 25% of confirmed cases in the state of Pará are exactly people in this age group, as can be seen in Fig. 4.

Figure 4. Percentage of confirmed cases and number of deaths by COVID-19 in the MRB, depending on the age group.



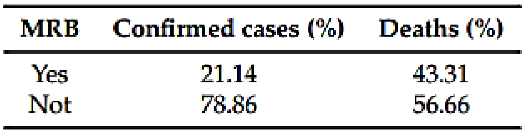
Source: Authors.

Likewise, more than 20% of the population of Pará are under the age of 20. However, less than 10% of confirmed cases are from people in this age group. The main characteristic of people in this age group is made up of students, and because of the social isolation measures to contain COVID-19, presential teaching activities were suspended, which may suggest that these people are less vulnerable to being infected.

The gardity of the disease is more noticeable of advanced age people. This is even more evident when the percentage of death is observed in people over 60 years of age. The elderly developed the most severe form of the disease when they tested positive for COVID-19, especially when they had a history of associated comorbidities such as: cardio vascular diseases, uncontrolled blood pressure, diebetes and kidney diseases, among others. People over 60 years old had a rate of 75% of registered deaths in the state of Pará, which clearly shows they are part of the risk group.

Another characteristic observed in this research is related to the place where infected people reside. The results presented in Table 2 show the percentage of confirmed cases and deaths by COVID-19, by patient's place of residence. It is noted that just over 20% of confirmed cases in the state of Pará, occurred in the MRB.

Table 2. Percentage of confirmed cases and deaths by COVID-19 in the state of Pará, Brazil by place of residence.



MRB	Confirmed cases (%)	Deaths (%)
Yes	21.14	43.31
Not	78.86	56.66

Source: Authors.

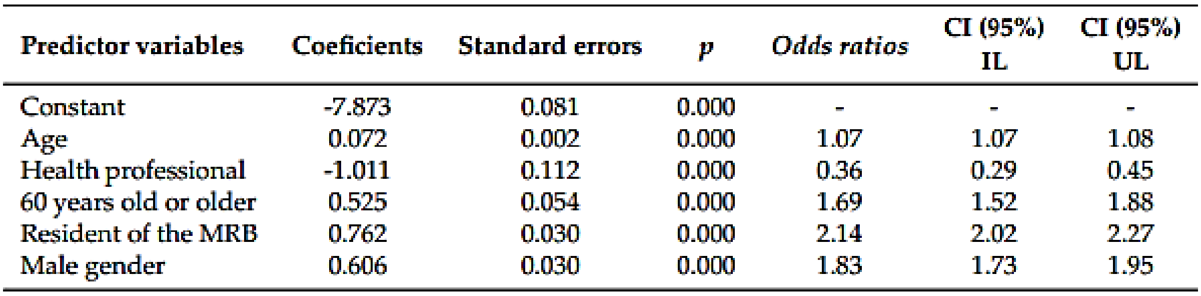
Due to the overcrowding of hospitals, the health system almost collapsed mainly in the capital (Belém, Brazil), where many patients did not receive adequate care, and some died. Another fact that we can verify is that more than 40% of the deaths in the state of Pará. It is worth mentioning that if there was a greater availability of clinical beds, both in the public and private network, many lives could have been preserved.

To estimate the death probability of a person due to COVID-19 in the metropolitan region of the city of Belém, a probabilistic model was used by means of multiple logistic regression (MLR). The application of MRL is used to obtain the statistical model that best fits the response variable.

The predictive variables of the model are: age (X_1), if you are a health professional (X_2), if you are 60 years old or older (X_3), if you are a resident of the MRB (X_4) and the male gender (X_5). The age variable is continuous and the other variables were coded into factors with two levels of classification: Yes or No, that is, they could assume two categories of coding. The variables X_1, X_2, X_3, X_4 and X_5 were coded in two categories: yes it receives the value one (1) and it does not present it receives the value zero (0).

Table 3 presents the parameter estimates for the model of the death probability by COVID-19 in the state of Pará (Yˆ 233) as well as the standard errors (SE), the values of the descriptive level (p), the odds ratios (OR), and the confidence intervals (CI) with the lower (IL) and upper (UL) limits.

Table 3. Estimates for the model of the probability.



Predictor variables	Coeficients	Standard errors	*p*	*Odds ratios*	CI (95%) IL	CI (95%) UL
Constant	-7.873	0.081	0.000	-	-	-
Age	0.072	0.002	0.000	1.07	1.07	1.08
Health professional	-1.011	0.112	0.000	0.36	0.29	0.45
60 years old or older	0.525	0.054	0.000	1.69	1.52	1.88
Resident of the MRB	0.762	0.030	0.000	2.14	2.02	2.27
Male gender	0.606	0.030	0.000	1.83	1.73	1.95

Source: Authors.

The statistical model obtained from the binary logistic regression for the probability of death by COVID-19 in the state of Pará (Yˆ), is given by Eq. 8:

$$\hat{Y} = \frac{exp(-7.873 + 0.072X_1 - 1.011X_2 + 0.525X_3 + 0.762X_4 + 0.606X_5)}{1 + exp(-7.873 + 0.072X_1 - 1.011X_2 + 0.525X_3 + 0.762X_4 + 0.606X_5)} \quad (8)$$

In Table 3, the positive coefficient ($\beta_1 = 0.072$) for X_1, suggests that deaths increase according to age. Thus, the odds ratio of 1.07 indicates that with each year older, a person diagnosed by COVID-19 has a greater than 7% probability of dying. This is only valid if the other variables are constant, which proves the severity of the disease for older people.

For variable X_2, the negative estimate ($\beta_2 = -1.011$) suggests that people who are health professionals are less likely to die. In addition, the odds ratio of 0.36 indicates that people diagnosed with COVID-19 who are not health professionals are almost 3 times more likely to die than health professionals, as long as the other variables are constant.

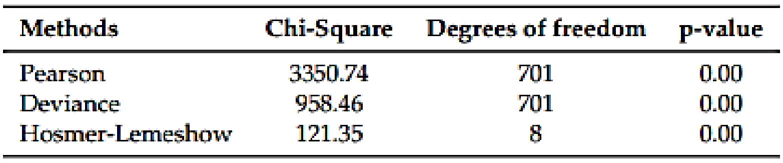
This probability for variable X_2 can be explained due to the fact that every health professional, in order to exercise their activities, needs to undergo a rigorous immunization process, taking all vaccines recommended by the Ministry of Health, of Brazil, through the Reference Centers in Special Immunobiologicals (Cries). The unhealthy environment in which they work must also be taken into account, which supports the development of various antibodies.

For variable X_3, the positive estimate ($\beta_3 = 0.525$) suggests that people aged 60 or over are more likely to die from COVID-19. In addition, the odds ratio of 1.69 indicates that people in this age group who test positive for coronavirus have a 69% greater chance of dying than people diagnosed under 60 years of age.

The positive coefficient (β_4 = 0.762) for variable X_4 suggests that people residing in MRB are more likely to die from COVID-19 than people diagnosed and who do not live in MRB. The odds ratio is 2.14, which means that, as long as the other characteristics are constant, a person residing in the MRB is more than twice as likely to die by COVID-19 than a person who lives in other cities. The high occupancy rates and even the overcrowding of clinical beds in the metropolitan region's health system may explain this high probability.

Finally, for variable X_5, the positive coefficient (β_5 = 0.606), suggests that the probability of death is higher in men than in women. The odds ratio of 1.83 indicates that a man has an 83% greater chance of dying from COVID-19 than a woman, keeping the other variables constant.

Table 4. Fit metrics with, the methods of Pearson, Deviance, and Hosmer-Lemeshow.

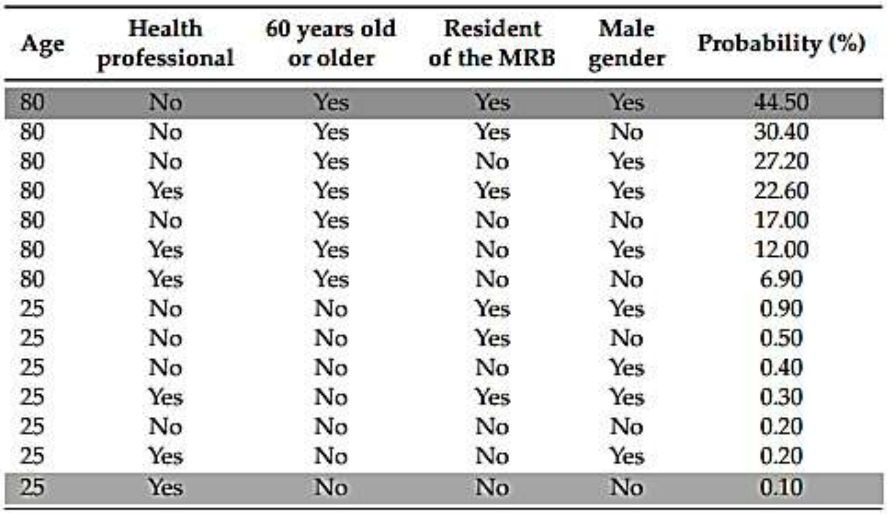


Methods	Chi-Square	Degrees of freedom	p-value
Pearson	3350.74	701	0.00
Deviance	958.46	701	0.00
Hosmer-Lemeshow	121.35	8	0.00

Source: Authors.

Table 5 shows the estimated probabilities for the adjusted multiple logistic regression model. After validation, we used it to estimate the probability of deaths from COVID-19 in the MRB. Where it was possible to observe, for example, that a man aged 80, who is not a health professional, resident of MRB, has a 44.50% probability of dying from COVID-19. On the other hand, a young man (25 years old), a health professional and not a resident in the MRB region, is less likely to die from COVID-19, with approximately 0.10% of chances.

Table 5. Death probabilities estimated from eq. 8.



Age	Health professional	60 years old or older	Resident of the MRB	Male gender	Probability (%)
80	No	Yes	Yes	Yes	44.50
80	No	Yes	Yes	No	30.40
80	No	Yes	No	Yes	27.20
80	Yes	Yes	Yes	Yes	22.60
80	No	Yes	No	No	17.00
80	Yes	Yes	No	Yes	12.00
80	Yes	Yes	No	No	6.90
25	No	No	Yes	Yes	0.90
25	No	No	Yes	No	0.50
25	No	No	No	Yes	0.40
25	Yes	No	Yes	Yes	0.30
25	No	No	No	No	0.20
25	Yes	No	No	Yes	0.20
25	Yes	No	No	No	0.10

Source: Authors.

5. Conclusions

This research proposed a probabilistic model based on a statistical model of non-linear regression analyzing the following variables (age, if you are a health professional, if you are 60 years old or older, if you are a resident of the MRB and the sex) to identify those that have the greatest impact on the number of infected and of deaths by COVID-19 in MRB. In order to carry out this research, data from all 151,849 people contaminated by COVID-19 in the state of Pará, Brazil were used up to that date, available from government agencies.

With the historical data available, it is observed that the peak of the disease in the State of Pará occurred in late May and early June, more precisely between May 21 and June 8, 2020. Although the percentage of men infected by the disease is lower than the women, it is clear that the majority of deaths were registered in men.

A similar situation can be observed when we analyze the age of infected people. Because people aged 30 to 49 years were the ones who most contracted the disease. However, people aged 50 and over account for the highest number of deaths recorded in the same period.

According to the probabilistic model of nonlinear regression proposed in this research, it is observed each year older between one person and another, diagnosed with COVID-19, has 7% more chances of dying when faced with the disease. Another, which proves the severity and the degree of lethality of the disease in the elderly. This can be explained because the elderly person, due to age, tends to develop several metabolic changes, in addition to having several comorbidities, which can worsen their health status.

Our work shows the importance of modeling the spread and deaths number, from different parameters, by COVID-19. The MLR model that we propose here, can help to assess the impacts of the disease by offering valuable predictions for making decisions, if proper restrictions and strong policies are implemented, before, during and after treating infected, to control the infection rates early from the spread of the disease.Thus, it was identified that men, the elderly and residents of the metropolitan region of Belém are people more likely to die from COVID-19, while health professionals are less likely.

References

Agresti, A. (1980). Generalized odds ratios for ordinal data. Biometrics, 36, 59–67. https://doi.org/10.2307/2530495

Allison, P. D. (2012). Logistic regression using SAS: Theory and application. In SAS Institute Inc, Second Edition, Cary, North Carolina, USA.

Bland, J. M., & Altman, D. G. (2000). The odds ratio. Bmj, (7247), 1468. https://doi.org/10.1136/bmj.320.7247.1468

Cabral, R. B. G., Chase, S. A. N., Ribeiro, R. C. M., Marques, G. T., Morais, E. C., de Jesus Zissou, A., de Almeida, J. F. S. (2020). On the evolution of new Coronavirus (SARS-CoV-2) in the state of Pará (Brazil), March – June 2020.

Candido, D. D. S., Watts, A., Abade, L., Kraemer, M. U., Pybus, O. G., Croda, J., Faria, N. R. (2020). Routes for COVID-19 importation in Brazil. Journal of Travel Medicine, taaa042.https://doi.org/10.1093/jtm/taaa042.

Cardoso, A. C. D., & Miranda, T. B. (2018). Invisibilidade social e produção do espaço subordinado em Belém (PA). Paisagem e Ambiente, (41), 85–107. https://doi.org/10.11606/issn.2359-5361.v0i41p85-107

Chambers, J. (2008). Software for data analysis: programming with R. In Springer Science and Business Media, Stanford, CA, USA.

Croda, J., Oliveira, W. K. D., Frutuoso, R. L., Mandetta, L. H., Baia-da-Silva, D. C., Brito-Sousa, J. D., & Lacerda, M. V. G. (2020). COVID-19 in Brazil: advantages of a socialized unified health system and preparation to contain cases. Revista da Sociedade Brasileira de Medicina Tropical, 53.

Dias, N. L., Silva, E. V. D., Pires, M. A., Chaves, D., Sanada, K. L., Fecury, A. A., Leal, S. D. (2020). Predição da propagação do SARS-CoV-2 no Estado do Amapá, Amazônia, Brasil, por modelagem matemática. Revista Científica Multidisciplinar Núcleo do Conhecimento, Volume 6(5), 73–95. (in portuguese)

Giancristofaro, R. A., & Salmaso, L. (2007). Model performance analysis and model validation in logistic regression. Statistica, 63(2), 375–396.

Hosmer Jr, D. W., Lemeshow, S., Sturdivant, R. X. (2013). Applied logistic regression. In John Wiley and Sons, Inc, Hoboken, New Jersey, 398.

IBGE (2010). Instituto Brasileiro de Geografia e Estatística (IBGE). In Retrieved from https://www.ibge.gov.br/estatisticas/sociais/populacao/9662-censo-demografico-2010.html ?edicao=9749&t=o-que-e, 2019.

Jesus, J. G., Sacchi, C., Claro, I., Salles, F., Manulli, E., Silva, D., Faria, N. R. (2020). First cases of Coronavirus Disease (COVID-19) in Brazil, South America (2020). United Kingdom: Virological, Retrieved from http://virological.org/t/first-cases-ofcoronavirus-disease-covid-19-in-brazil-south-america-2-genomes-3rd.

Kermack, W. O., & McKendrick, A. G. A (1927). Contribution to the mathematical theory of epidemics. Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character, 115(772), 346 700–721.

Kounev, S., Lange, K. D., von Kistowski, J. (2020). Review of basic probability and statistics. In Systems Benchmarking, Springer, Cham, 23–44).

Lewis-Beck, C., and Lewis-Beck, M. (1984). Applied regression: An introduction. InSage publications, 22, 384.

Martins, L. K., Carvalho, A. R. D. S., Oliveira, J. L. C. D., Santos, R. P. D., Lordani, T. V. A. (2020). Quality of life and perception of health status among hospitalized individuals, Escola Anna Nery, 24(4), e20200065, 405 https://doi.org/10.1590/2177-9465-ean-2020-0065

Menard, S. (2002). Applied logistic regression analysis. In Sage publication, 106, Thosand Oak, London, New Delhi.

Mendes, L. A. S. (2018). A Geografia-Histórica da região metropolitana de Belém. Revista Espacialidades, 14(01), 10–39.

Neter, J., Wasserman, W., Kutner, M. H.(1983). Applied linear statistical models. In Richard D. Irwin. Inc., Homewood, IL, 842.

Oliveira, W. K. D., Duarte, E., França, G. V. A. D., Garcia, L. P. (2020). How Brazil can stop COVID-19. Epidemiologia e Serviços de Saúde, 29(2), e2020044. https://doi.org/10.5123/S1679-49742020000200023. (in portuguese).

Oliveira Ribeiro, W. (2016). Entre a metrópole e a cidade média: a complexidade das interações espaciais e 351 das dinâmicas de centralidade da cidade de Castanhal no nordeste paraense. GEOUSP Espaço e Tempo(Online), 20(1), 115–129.

Petropoulos, F., & Makridakis, S., Forecasting (2020). The novel coronavirus COVID-19. PloS one, 15(3), e0231236. https://doi.org/10.1371/journal.pone.0231236

Pregibon, D. (1981). Logistic regression diagnostics. The Annals of Statistics, 9(4), 705–724.

Priyadarshan, P. M. (2019). Basic Statistics. In PLANT BREEDING: Classical to Modern, Springer, Singapore, 131–169.

Pulkstenis, E., and Robinson, T. J. (2002). Two goodness-of-fit tests for logistic regression models with continuous covariates. Statistics in medicine, 21(1), 79–93. https://doi.org/10.1002/sim.943

Qin, X., Chen, C., Yam, K. C., Huang, M., Ju, D. (2020). The double-edged sword of leader humility: Investigating when and why leader humility promotes versus inhibits subordinate deviance. Journal of Applied Psychology, 105(7), 693–712. https://doi.org/10.1037/apl0000456

Romano, J. P., & Wolf, M. (2005). Stepwise multiple testing as formalized data snooping. Econometrica, 73(4), 1237–1282. https://doi.org/10.1111/j.1468-0262.2005.00615.x

Sarkar, S. K. (2002). Some results on false discovery rate in stepwise multiple testing procedures. Annals of statistics, 30(1), 239–257. Retrieved from https://www.jstor.org/stable/2700010

Silva, P. C., Batista, P. V., Lima, H. S., Alves, M. A., Guimarães, F. G., Silva, R. C. (2020). COVID-ABS: An agent-based 348 model of COVID-19 epidemic to simulate health and economic effects of social distancing interventions. Chaos, Solitons & Fractals, 110088. https://doi.org/10.1016/j.chaos.2020.110088

Szumilas, M. (2010). Explaining odds ratios. Journal of the Canadian academy of child and adolescent psychiatry, 19(3), 227–229.

Xie, X. J., Pendergast, J., Clarke, W. (2002). Increasing the power: A practical approach to goodness-of-fit test for logistic regression models with continuous predictors. Computational Statistics and Data Analysis, 52(5), 2703–2713. https://doi.org/10.1016/j.csda.2007.09.027

Werneck, G. L., and Carvalho, M. S (2020). A pandemia de COVID-19 no Brasil: crônica de uma crise sanitária anunciada. Caderno de Saúde Pública, 36(5), https://doi.org/10.1590/0102-311X00068820

Zhou, P., Yang, X. L., Wang, X. G., Hu, B., Zhang, L., Zhang, W., ... Chen, H. D. A (2020). Pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature, 579(7798), 270–273. https://doi.org/10.1038/s41586-020-2012-7

Percentage of contribution of each author in the manuscript

Cássio Pinho dos Reis - 20%

Herson Oliveira da Rocha - 15%

Nayara de Araújo Muzili Reis - 15%

Sávio Pinho dos Reis Reis - 15%

Gustavo Nogueira Dias - 10%

Gilberto Emanoel Reis Vogado - 10%

Vanessa Mayara Souza Pamplona - 10%

Washington Luiz Pedrosa da Silva Junior - 5%

Abbreviations

The following abbreviations are used in this manuscript:
CI Confidence Intervals
COE Emergency Operations Center
ESPII Public Health Emergency of International Importance
IAL Institute Adolfo Lutz
MLE Maximum Likelihood Estimation
MLR Multiple Logistic Regression
MS Ministry of Health OR Odds Ratios
RMB Metropolitan Region of Belém
SE Standard Errors
SEIR Susceptible-Exposed-Infected-Recovered
SESPA State Department of Health of Pará
SIR Susceptible-Infected-Recovered
SVS Health Surveillance Secretariat
WHO World Health Organization