UNIVERSIDADE ESTADUAL DA PARAÍBA
CAMPUS I - CAMPINA GRANDE
CENTRO DE CIÊNCIAS E TECNOLOGIA
DEPARTAMENTO DE ESTATÍSTICA
CURSO DE GRADUAÇÃO EM ESTATÍSTICA

FÁTIMA AGATHA BERTULINO LIMA

AJUSTES DE MODELOS LINEARES GENERALIZADOS PARA OS DADOS NOTIFICADOS POR COVID-19 NO ESTADO DA PARAÍBA

CAMPINA GRANDE -PB
2023

FÁTIMA AGATHA BERTULINO LIMA

AJUSTES DE MODELOS LINEARES GENERALIZADOS PARA OS DADOS NOTIFICADOS POR COVID-19 NO ESTADO DA PARAÍBA

Trabalho de Conclusão de Curso apresentado ao curso de Bacharelado em Estatística do Departamento de Estatística do Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba, como requisito parcial à obtenção do título de bacharel em Estatística.

Área de concentração: Estatística

Orientador: Prof.Dr. Tiago Almeida de Oliveira.

CAMPINA GRANDE -PB
2023

L732a Lima, Fatima Agatha Bertulino.

Ajustes de modelos lineares generalizados para os dados notificados por Covid-19 no estado da Paraíba [manuscrito] / Fatima Agatha Bertulino Lima. - 2023.

32 p. : il. colorido.

Digitado.

Trabalho de Conclusão de Curso (Graduação em Estatística) - Universidade Estadual da Paraíba, Centro de Ciências e Tecnologia, 2023.

"Orientação : Prof. Dr. Tiago Almeida de Oliveira, Coordenação do Curso de Estatística - CCT. "

1. Modelos lineares generalizados. 2. Variáveis socioeconômicas. 3. Macrorregiões de saúde. I. Título

21. ed. CDD 519.5

Elaborada por Geovani S. de Oliveira - CRB - 15/1009

Biblioteca Central BC/UEPB

FÁTIMA AGATHA BERTULINO LIMA

AJUSTES DE MODELOS LINEARES GENERALIZADOS PARA OS DADOS NOTIFICADOS POR COVID-19 NO ESTADO DA PARAÍBA

Trabalho de Conclusão de Curso apresentado ao curso de Bacharelado em Estatística do Departamento de Estatística do Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba, como requisito parcial à obtenção do título de bacharel em Estatística.

Área de concentração: Estatística

Aprovado em: 27/11/2023

BANCA EXAMINADORA

Prof. Dr. Tiago Almeida de Oliveira (Orientador)
Universidade Estadual da Paraíba (UEPB)

Prof. Dr. Silvio Fernando Alves Xavier Junior
Universidade Estadual da Paraíba (UEPB)

Prof. Me. Cleanderson Romualdo Fidelis
Universidade Estadual da Paraíba (UEPB)

RESUMO

A modelagem de dados de saúde frequentemente desafia os pressupostos da regressão clássica Gaussiana, que pressupõe que a variável resposta seja simétrica e homocedástica. Para contornar essas limitações, emprega-se a abordagem dos Modelos Lineares Generalizados (MLGs), caracterizada por sua flexibilidade ao permitir diferentes distribuições para a variável resposta. Este estudo aplicou ajustes utilizando as distribuições gama e normal inversa para analisar os casos notificados por Covid-19 nas três macrorregiões de saúde na Paraíba. O objetivo foi compreender a influência de variáveis socioeconômicas e sociodemográficas na quantidade de casos notificados. Na Macrorregião 1, identificou-se que o índice de Gini, a educação e a razão de dependência exercem impacto significativo na distribuição de casos. Na Macrorregião 2, variáveis como índice de Gini, a renda, a longevidade, a educação e a razão de dependência contribuíram para o modelo final. Na Macrorregião 3, o índice de Gini, a renda, a educação e a razão de dependência foram variáveis significativas. Esta pesquisa oferece contribuições significativas para o entendimento dos determinantes socioeconômicos na propagação da Covid-19 na Paraíba, informando estratégias de intervenção e políticas públicas direcionadas a diferentes realidades populacionais. A utilização da abordagem dos MLGs com distribuições específicas evidenciou-se como uma ferramenta robusta, superando as limitações dos modelos clássicos. Os resultados destacam a importância de considerar a heterogeneidade regional na formulação de políticas de saúde, pois contribuem não apenas para a área acadêmica, mas também para orientar ações efetivas no enfrentamento de futuras ameaças coletivas.

Palavras-chaves: modelos lineares generalizados; variáveis socioeconômicas; macrorregiões de saúde.

ABSTRACT

The modeling of health data often challenges the assumptions of classical Gaussian regression, which presupposes that the response variable is symmetric and homoscedastic. To overcome these limitations, the approach of Generalized Linear Models (GLMs) is employed, characterized by its flexibility in allowing different distributions for the response variable. This study applied adjustments using gamma and inverse normal distributions to analyze Covid-19 cases reported in the three health macro-regions in Paraíba, Brazil. The aim was to understand the influence of socioeconomic and sociodemographic variables on the reported case count. In Macro-region 1, it was identified that the Gini index, education, and dependency ratio have a significant impact on the distribution of cases. In Macro-region 2, variables such as the Gini index, income, longevity, education, and dependency ratio contributed to the final model. In Macro-region 3, the Gini index, income, education, and dependency ratio were significant variables. This research provides significant contributions to understanding the socioeconomic determinants in the spread of Covid-19 in Paraíba, informing intervention strategies and public policies tailored to different population realities. The use of GLMs with specific distributions proved to be a robust tool, overcoming the limitations of classical models. The results emphasize the importance of considering regional heterogeneity in formulating health policies, contributing not only to the academic field but also guiding effective actions in addressing future collective threats.

Keywords: generalized linear models; socioeconomic variables; health macro-regions.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Conforme Breno et. al. (2020) o coronavírus, inicialmente identificado em 1937, ganhou destaque em 2002-2003 ao causar a Síndrome Respiratória Aguda Grave (SARS). Embora tenha sido controlado rapidamente, seu sucessor, o SARS-CoV-2(síndrome respiratória aguda grave do tipo 2), surgiu em 2019, espalhando-se globalmente. Menos letal que seus predecessores, o novo vírus, responsável pela Covid-19, apresenta maior potencial de disseminação, a epidemiologia da Covid-19 ainda é pouco compreendida, exigindo ações de saúde pública para mitigar a morbimortalidade e conter a propagação.

Segundo Campos et. al. (2020) em dezembro de 2019, o primeiro caso oficial de pneumonia, inicialmente desconhecida, foi registrado em Wuhan, China, despertando a atenção global. A síndrome respiratória aguda grave (SRAG), posteriormente denominada Cocid-19 pela OMS, foi caracterizada como pandemia em março de 2020. A transmissibilidade do vírus (R0) foi inicialmente registrada em 2,2, indicando uma média de duas pessoas infectadas por cada pessoa infectada. A Covid-19, causada pelo SARS-CoV-2, apresenta rápida disseminação, com uma variação de R0 estimado de 1,95 a 6,5. A maioria dos casos (70-80%) é assintomática ou leve, mas 20% podem desenvolver formas graves, exigindo cuidados hospitalares. A gravidade está associada à idade avançada e comorbidades. Até maio de 2020, globalmente, foram registrados cerca de 6 milhões de casos e 365 mil mortes, enquanto o Brasil tinha aproximadamente 470 mil casos e 30 mil mortes, com taxas de letalidade de 6%.

De acordo com o Ministério de Saúde (2022), quanto ao modo de transmissão, o SARS-CoV-2 é principalmente disseminado por contato, gotículas e partículas ou aerossóis respiratórios. A infecção ocorre através da inalação de gotículas ou partículas de aerossol, contato direto com membranas mucosas e toque em superfícies contaminadas. No Brasil, a vigilância epidemiológica desses vírus é conduzida por meio de uma Rede de Vigilância Sentinela de síndrome gripal (SG) e Vigilância de síndrome respiratória aguda grave (SRAG), visando traçar medidas de prevenção e controle com base no perfil epidemiológico e nos vírus circulantes.

Diante dessa pandemia, segundo Lacerda (2020) a estatística vem ganhando muito destaque devido as políticas de enfrentamento que vem sendo baseadas em estudos epidemiológicos, onde são utilizados os modelos estatísticos, como, por exemplo para estimar o número de casos em diferentes cenários e localizações, assim, auxiliando os tomadores de decisão a determinar número de vagas em UTI.

Neste estudo, a variável resposta é uma variável de contagem, o número de casos notificados por Covid-19, em que tem-se o objetivo de analisar como esse número é supostamente influenciado pelas variáveis com informações socioeconômicas e sociodemográficas, as variáveis explicativas. Segundo Dobson e Barnett (2008), o modelo utilizado frequentemente na análise desse tipo de dados é o modelo de regressão linear gaussiana. Como o número de casos notificados é uma contagem, constitui a ocorrência de um dado número de eventos durante um intervalo de tempo ou espaço, porém nem sempre as suposições de normalidade e homoscedasticidade

dos erros, característicos desse modelo são satisfeitas (Conceição, 2001).

Então tem-se a possibilidade de usar modelos estatísticos análogos aos modelos de regressão gaussiana, nas situações em que a variável resposta não satisfaz as características do modelo linear. Nestes casos, geralmente utilizam-se as classes de modelos que oferecem uma poderosa alternativa para a transformação de dados, chamadas de modelos lineares generalizados (MLGs) ou aditivos (GAMs), que são uma extensão dos modelos de regressão linear ou não paramétricos, permitindo que possamos trabalhar com a variável de interesse mesmo que ela siga outros tipos de distribuições (Schmidt, 2003).

Diante do exposto, esta pesquisa tem o objetivo de analisar possíveis correlações entre os casos notificados por Covid-19 e variáveis socioeconômicas e sociodemográficas, assim, identificar o modelo mais indicado para as amostras, divididas por macrorregiões de saúde do estado da Paraíba, estudando sobre a influência das variáveis explicativas no aumento ou diminuição da média do número de casos notificados, através de técnicas estatísticas para ajustes de modelos lineares generalizados (MLGs).

2 MATERIAL E MÉTODOS

2.1 Material

Os dados dessa pesquisa são provenientes do Ministério da Saúde, Secretaria de Vigilância em Saúde (SVS): Guia de Vigilância Epidemiológica do Covid-19, a faixa temporal de evolução dos casos nesta análise compreendeu o período de janeiro de 2021 a abril de 2022, que estão disponíveis no site Brasil.io[1]. Para os dados socioeconômicos e sociodemográficos, foram obtidos através do site Atlas[2]. Todas as análises estatísticas foram realizadas com a linguagem de programação Software R (versão 4.2.2) e serão ilustradas na forma de tabelas e gráficos.

2.1.1 Macrorregiões

Segundo Brandão (2012), macrorregiões são compostas por um determinado número de municípios agrupados de acordo com as características demográficas, sócio-econômicas, sanitárias, epidemiológicas, de acessibilidade e de oferta de serviços de saúde, em que os municípios sedes são responsáveis por absorver as maiores demandas provenientes das cidades vizinhas.

O Centro Formador de Recursos Humanos da Paraíba (CEFOR) (2019) destaca que a subdivisão em macrorregionais visa proporcionar à população acesso de qualidade e ações de saúde contínuas de qualquer ponto da rede. Essa estratégia busca facilitar o acesso, reduzir custos e deslocamentos, concentrando esforços na promoção da saúde e prevenção de processos de doença.

Na Tabela 1, apresenta-se a divisão de macrorregiões aprovada em 2018 pela Secretaria de Saúde do Estado da Paraíba.

Tabela 1 – Macrorregiões da saúde na Paraíba

Macrorregião	Município sede	Regiões da Saúde	População
1ª	João Pessoa	1ª,2ª,12ª e 14ª	1.952.127
2ª	Campina Grande	3ª,4ª,5ª,15ª e 16ª	1.127.117
3ª	Patos (Sertão), Sousa (Alto Sertão)	6ª,7ª,8ª,9ª,10ª,11ª e 13ª	946.314

Fonte: Elaborada pela autora, 2023.

2.1.2 Variáveis

Neste estudo tem-se a análise das variáveis explicativas descritas na Tabela 2, com o objetivo de investigar o impacto de características socioeconômicas e sociodemográficas na variável resposta em estudo, os casos notificados por Covid-19.

[1] `https://brasil.io/covid19/PB/`
[2] `http://www.atlasbrasil.org.br./`

Tabela 2 – Codificação das variáveis explicativas e da variável resposta

Variável	Codificação
casos	Quantidade de casos notificados (2022)
pop	Quantidade da população (2016)
gini	Índice de Gini (2010)
renda	IDH de renda (2010)
long	IDH de longevidade (2010)
edu	IDH de educação (2010)
rdep	Razão de dependência (2010)
pib	Produto Interno Bruto per capita (2016)

Fonte: Elaborada pela autora, 2023.

Segundo Wolffenbüttel (2004), o Índice de Gini, criado pelo matemático italiano Conrado Gini, é um instrumento para medir o grau de concentração de renda em determinado grupo. Ele aponta a diferença entre os rendimentos dos mais pobres e dos mais ricos. Numericamente, varia de zero a um, sendo, o valor 0 (zero) representa a situação de igualdade, ou seja, todos têm a mesma renda, enquanto o 1 (um) representa que só uma pessoa detém toda a riqueza.

Para Sousa (2023), o Índice de Desenvolvimento Humano (IDH) é uma unidade de medida utilizada para aferir o grau de desenvolvimento de uma determinada sociedade nos quesitos de educação, saúde e renda. O IDH é uma referência numérica que varia entre 0 e 1. Quanto mais próximo de zero, menor é o indicador para os quesitos de saúde (longevidade), educação e renda.

De acordo com o Ministério de Saúde, a Razão de dependência é a razão entre a população potencialmente inativa (0 a 14 anos e 65 anos ou mais de idade) e a população potencialmente ativa (15 a 64 anos de idade), na data de referência do Censo Demográfico (DATASUS, 2023).

Conforme o site do Instituto Brasileiro de Geografia e Estatística (IBGE), o PIB é a soma de todos os bens e serviços finais produzidos por um país, estado ou cidade, geralmente em um ano. Os bens e serviços finais que compõem o PIB são medidos no preço em que chegam ao consumidor. Dessa forma, levam em consideração também os impostos sobre os produtos comercializados (IBGE, 2023).

2.2 Metodologia

Este estudo utiliza métodos estatísticos, inicialmente com a análise descritiva para cada macrorregião de saúde do estado da Paraíba. Para modelar a variável resposta, a quantidade de casos notificados por Covid-19, foi associada a apenas uma variável explicativa utilizando regressão simples, posteriormente acrescentando variáveis explicativas para a abordagem por regressão múltipla, em que não atendeu aos pressupostos da regressão linear gaussiana, portanto, foram realizados ajustes por modelos lineares generalizados para a variável resposta que se trata de dados numéricos.

2.2.1 Família exponencial uniparamétrica

Segundo Cordeiro e Demétrio (2008), a família exponencial uniparamétrica é caracterizada por uma função (de probabilidade ou densidade) especificada na forma

$$f(x;\theta) = h(x)\exp[\eta(\theta)t(x) - b(\theta)], \quad (2.1)$$

em que as funções $\eta(\theta)$, $b(\theta)$, $t(x)$ e $h(x)$ têm valores em subconjuntos dos reais, várias distribuições importantes podem ser expressas nessa forma, tais como: poisson, binomial, normal, gama, sendo a normal em gama com suposição de que um dos parâmetros é conhecido. Na parametrização (2.1) θ é denominado de parâmetro canônico. O logaritmo da função de verossimilhança corresponde a uma única observação no modelo. De acordo com Nelder e Wedderburn (1972), os MLGs são definidos por uma distribuição de probabilidade, membro da família exponencial de distribuições, e são formados pelas seguintes componentes:

- **Componente aleatório:** n variáveis explicativas $X_1, X_2, ..., X_n$, de uma variável resposta que segue uma distribuição da família exponencial com valor esperado $E(X_i) = \mu$;

- **Componente sistemático:** compõe uma estrutura linear para o modelo de regressão $\eta = X^T\beta$, chamado de preditor linear onde $\mathbf{X^T} = (\mathbf{x_{i1}}, \mathbf{x_{i1}}, ..., \mathbf{x_{ip}})^\mathbf{T}$, em que $i = 1, 2, ..., n$ são as variáveis explicativas;

- **Função de ligação:** uma função monótona e diferenciável g, capaz de conectar as componentes aleatória e sistemática, ou seja, relaciona a média da variável resposta (μ) à estrutura linear, definida nos MLG por $g(\mu) = \eta$, onde $\eta = X^T\beta$, com o coeficiente de regressão $\beta = (\beta_1, \beta_2, ..., \beta_n)$ representando o vetor de parâmetros a ser estimado.

Segundo Myers e Montgomery (2002), a escolha da função de ligação em modelo linear generalizado pode ser considerada como o equivalente à escolha de uma transformação da variável resposta Y_i em um modelo linear de regressão, contudo, é relevante explicar que a função de ligação transforma μ_i, a média de Y_i, e não a variável resposta.

Na Tabela 3 identifica-se as ligações canônicas de acordo com a distribuição e a natureza de dados em que é utilizada

Tabela 3 – Ligações canônicas para os MLG

Distribuição	Ligação canônica	Tipos de dados
normal	$\eta = \mu$	contínuo
poisson	$\eta = \ln\mu$	contagem
binomial	$\eta = ln(\pi/(1-\pi))$	proporção
gama	$\eta = 1/\mu$	contínuo assimétrico
normal inversa	$\eta = 1/\mu^2$	contínuo assimétrico

Fonte: Elaborada pela autora, 2023.

2.2.2 Modelo de Regressão Linear Simples

Segundo Hoffman et. al. (2016), dados n pares de valores de duas variáveis X_i, Y_i (com i=1,2,...,n), se admitirmos que Y_i é função linear de X_i, podemos estabelecer uma regressão linear cujo modelo é:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \tag{2.2}$$

em que $i = 1, 2, ..., n$. Onde os β são parâmetros e X é a variável explicativa e Y é a variável dependente. O modelo é chamado de linear, por impor que a função da média (μ) é uma função linear do coeficiente (β_0, β_1) que são os elementos do vetor de parâmetros β e simples porque tal função envolve apenas uma variável explicativa X. Para se fazer inferência estatística, ao obtermos nosso modelo de regressão é preciso fazer testes de pressupostos afim de descobrir se o modelo realmente esta adequado aos nossos dados, sendo eles:

- Linearidade: a relação entre as variáveis independentes e dependentes possa ser expressa como uma função linear.
- Homoscedasticidade (ou Homogeneidade de Variância): os termos de erro tem variância constante, independente dos valores das variáveis explicativas
- Independência de erros: os erros nas variáveis preditoras não devem estar correlacionados.
- Não multicolinearidade: as variáveis preditoras não podem ser próximas de uma correlação perfeita.
- Baixa exogeneidade: os valores das variáveis preditoras não estão contaminados com erros de medida.
- Os erros seguem distribuição normal, implicando que a variável resposta Y também tenha distribuição normal

2.2.3 Modelo de Regressão Linear Múltiplo

De acordo com Hoffman et. al. (2016), temos uma regressão linear múltipla quando admitimos que o valor da variável dependente é função linear de duas ou mais variáveis explanatórias. O modelo estatístico de uma regressão linear múltipla com k variáveis explanatórias é:

$$Y_i = \beta_0 + X_{i1}\beta_1 + X_{i2}\beta_2 + ... + \varepsilon_i, \tag{2.3}$$

em que $i = 1, 2, ..., n$. Nesse modelo a variável dependente Y_i é função linear das variáveis explicativas $X_{ik}, (i = 1, ..., k)$, mantendo os pressupostos da linear simples, a modificação é que temos mais variáveis para adicionar ao modelo e analisar as influencias na variável resposta. A estimação dos parâmetros é feita através da maximização por Máxima Verossimilhança (MV) ou Método dos Mínimos Quadrados (MMQ), pode-se encontrar mais detalhes em Clarice et. al. (1999). Segundo Rodrigues (2012), as variáveis independentes, designadas como variáveis

explicativas, desempenham o papel de explicar a variação em Y_i. Na regressão linear múltipla, presume-se uma relação linear entre uma variável dependente e as variáveis independentes (preditoras).

2.2.4 Modelo Linear Generalizado

Os Modelos Lineares Generalizados foram propostos por Nelder e Wedderburn (1972) como uma extensão do modelo linear normal, diferente dos modelos linear simples e múltiplo que possuem o pressuposto de normalidade e os erros homocedásticos, segundo Dobson e Barnett (2008) os MLGs são modelos que possibilitam o ajuste aos erros utilizando outras famílias de distribuições como exponencial, poisson, gama, assim permitindo a modelagem de variáveis respostas em forma de contagem, contínuas simétricas ou assimétricas, binárias e categóricas. Os modelos generalizados são usados para resíduos heterocedásticos, em que a variância varia em função da média, considerando que o parâmetro de dispersão é constante em todas observações, essa propriedade se dá pelo fato que os parâmetros dos modelos MLGs são estimados através do mínimos quadrados ponderados, assim desconsideramos a suposição de homocesticidade dos resíduos, também fazem o uso de funções de ligação, relacionando a média da variável resposta à combinação linear das variáveis explicativas (Malveira, 2018). Segundo Barros (2016) para os MLGs não se define uma distribuição para o erro como ocorre no modelo de regressão normal linear, mas sim para a variável resposta Y_i.

Neste estudo, realizou-se o ajuste dos dados por modelos lineares generalizados, empregando as distribuições gama, binomial negativa e normal inversa, devido a melhor adequação da variável resposta. Deste modo, de acordo com Paula (2013), Turkman (2000), tem-se as propriedades seguidamente para as distribuições.

2.2.4.1 Gama

Admitindo- se que a variável resposta Y_i é aleatória com distribuição gama de média μ e coeficiente de variação $\phi^{\frac{1}{2}}$ denota-se $Y_i \sim G(\mu_i, \phi)$;

- A esperança é dada por μ;
- A variância é dada por μ^2, consequentemente a dispersão aumenta à medida que a média aumenta;
- Tem-se que o valor esperado de μ_i está relacionado com o preditor linear $\eta_i = X^T\beta$ através da relação $\mu_i = exp(\eta_i)$.
- A função densidade de probabilidade é expressa por

$$f(y;\mu,\phi) = \frac{1}{\Gamma(\phi)}\left(\frac{y}{\mu}\right)^{\phi-1} e^{-\frac{y}{\mu}} \tag{2.4}$$

2.2.4.2 *Binomial Negativa*

Admitindo- se que a variável resposta Y_i é aleatória com distribuição binomial negativa de média μ e parâmetro de forma ϕ denota-se $Y_i \sim BN(\mu_i, \phi)$;

- A esperança é dada por μ;
- A variância é dada por $\mu + \mu^2/\phi$
- Tem-se que o valor esperado de μ_i está relacionado com o preditor linear $\eta_i = X^T\beta$ através da relação $\mu_i = \log(\eta_i)$.
- A função de probabilidade massiva é expressa por

$$P(Y = y; \mu, \phi) = \binom{y + \phi - 1}{y} \left(\frac{\mu}{\mu + \phi}\right)^{\phi} \left(\frac{\phi}{\mu + \phi}\right)^{y} \tag{2.5}$$

2.2.4.3 *Normal Inversa*

Admitindo- se que a variável resposta Y_i é aleatória com distribuição normal inversa de média μ e parâmetro de forma ϕ denota-se $Y_i \sim NI(\mu_i, \phi)$;

- A esperança é dada por μ;
- A variância é dada por μ^3;
- Considera-se que o valor esperado de μ_i está relacionado com o preditor linear $\eta_i = X^T\beta$ através da relação $\mu_i = (\exp(\eta_i))^{\frac{1}{2}}$.
- A função densidade de probabilidade é expressa por

$$f(y; \mu, \phi) = \sqrt{\frac{\phi}{2\pi y^3}} e^{-\frac{\phi}{2y}\left(\frac{y-\mu}{\mu y}\right)} \tag{2.6}$$

2.2.5 Métodos de estimação

O método de estimação para o vetor de parâmetros β, proposto por Nelder e Wedderburn (1972), utilizam o método de máxima verossimilhança. A log-verossimilhança da família exponencial será dada por

$$\log_e(L(\theta, \phi, y)) = \sum_{i=1}^{n} \left(\frac{y_i\theta_i + b(\theta_i)}{\phi} + c(y_i, \phi)\right) \tag{2.7}$$

onde $\theta = (\theta_1, \theta_2, ..., \theta_n)$ e $y = (y_1, y_2, ..., y_n)$. De acordo com Lucambio (2022), suponha que que um modelo linear generalizado use a função de ligação $g()$, de modo que,

$$g(\mu_i) = \eta_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_k X_{ik} \tag{2.8}$$

em que $i = 1, 2, ..., n$. O modelo, portanto expressa os valores esperados das n observações em termos de um número muito menor de parâmetros de regressão. Para obter equações de estimação para os parâmetros de regressão, temos que diferenciar o logaritmo da verossimilhança em relação a cada coeficiente por sua vez.

Segundo Lucambio (2022), como os valores de μ_i e η_i são desconhecidos, que, de fato, dependem dos coeficientes de regressão que desejamos estimar, assim o argumento é essencialmente circular. Esta observação sugeriu a Nelder e Wedderburn (1972) a possibilidade de estimar os modelos lineares generalizados por mínimos quadrados ponderados iterativos (MQPI), habilmente transformando a circularidade em um procedimento iterativo:

1. Calcular as estimativas iniciais de $\hat{\mu}_i$ e $\hat{\eta}_i = g(\hat{\mu}_i)$;

2. Em cada iteração l, calcule a variável resposta de trabalho Z usando os valores de $\hat{\mu}$ e $\hat{\eta}$ da iteração anterior, junto com os pesos W

$$Z_i^{(l-1)} = \eta_i^{(l-1)} + (y_i + \mu_i^{(l-1)}) g'(\mu_i^{(l-1)}), \tag{2.9}$$

$$W_i^{(l-1)} = \frac{1}{(g'(\mu_i^{(l-1)})) + (\mu_i^{(l-1)})}; \tag{2.10}$$

3. Ajustar um modelo de regressão de mínimos quadrados ponderados de $Z_i^{(l-1)}$ nos Xs usando $W_i^{(l-1)}$ como pesos;

4. Repetir as etapas 2 e 3 até que os coeficientes de regressão se estabilizem, ponto em que $\hat{\beta}$ converge para as estimativas de máxima verossimilhança dos βs.

2.2.6 Qualidade do ajuste

Para Paula (2013), a qualidade do ajuste de um MLG é avaliada através da função desvio

$$D(y, \hat{\mu}) = 2 \sum_{i=1}^{n} y_i(\theta_i - \hat{\theta}_i) = (b(\hat{\theta}_i - b(\theta_i)) \tag{2.11}$$

que é uma distância entre o logaritmo da função de verossimilhança do modelo saturado (com n parâmetros) e do modelo sob investigação (com $p = n$ parâmetros) avaliado na estimativa de máxima verossimilhança $\hat{\beta}$. Um valor pequeno para a função desvio indica que, para um número menor de parâmetros, obtemos um ajuste tão bom quanto o ajuste com modelo saturado.

2.2.7 Seleção de variáveis

Segundo Alves et. al. (2013), o método *stepwise* é usado para selecionar quais variáveis independentes mais influenciam a variável resposta podendo, assim, diminuir o número de variáveis que compõe o modelo de regressão. O método é feito de forma iterativa, adicionando (passo *forward*) e removendo variáveis (passo *backward*), a partir de um critério de seleção, um

dos critérios de seleção mais usados é o teste F, mas também pode ser feito com o coeficiente correlação linear múltipla, erro quadrático total, critério de informação de Akaike Hoking (1974).

Iwata e Sandoval (2018) descreveu os processos para as seleções de variáveis da seguintes forma:

Seleção *forward*

1. Comece com uma regressão com apenas o intercepto β_0;
2. Para as demais variáveis candidatas, escolha aquela cuja inclusão implica em maior aumento de R^2;
3. Se essa nova adição foi estatisticamente significante, mantenha a variável; caso contrário, retire a variável, volte ao modelo anterior, e pare o algoritmo;
4. Repita os passos 2 e 3 até que a adição de qualquer nova variável não seja estatisticamente significante (a um nível de significância pré-especificado).

Seleção *backward*

1. Comece com uma regressão com todas as variáveis candidatas;
2. Se houver alguma variável cujo coeficiente é estatisticamente não significativo, elimine a variável que tenha menor nível de significância no modelo (maior p-valor); caso contrário, esse é o modelo final;
3. Repita o passo 2 até atingir um modelo no qual todas as variáveis são estatisticamente significantes (a um nível de significância pré-definido).

Seleção *stepwise*

1. Trata-se de uma combinação das seleções do tipo *forward* e *backwards*;
2. Os passos *forward* e *backwards* são intercalados, de forma a adicionarmos variáveis que sejam significativas e retirarmos variáveis que não sejam estatisticamente significativas;
3. O algoritmo para quando não for mais possível adicionar variáveis novas que sejam estatisticamente significantes, ou retirar variáveis incluídas que forem estatisticamente não significantes.

2.2.8 Técnica de disgnóstico: Multicolinearidade

Segundo Gori (2017), há multicolinearidade em um modelo de regressão múltipla quando duas ou mais variáveis independentes são fortemente relacionadas linearmente entre si. A existência de uma colinearidade exata entre duas ou mais variáveis independentes torna impossível a obtenção dos coeficientes dos parâmetros por MQO. Uma maneira de medir a multicolinearidade é o fator de inflação da variância (VIF), que avalia o quanto a variância de um coeficiente de

regressão estimado aumenta se as suas preditoras estiverem correlacionadas. Se nenhum fator estiver correlacionado, os VIFs serão todos 1, é recomendado que valores de VIF maiores do que 5 podem causar sérios problemas na estimação dos coeficientes de regressão (Draper e Smith, 1998).

2.2.9 Seleção de Modelos Generalizados

Para a escolha do modelo com a distribuição e função de ligação mais adequada ao dados, utiliza-se o *Akaike Information Criterion*, AIC, que de acordo com Akaike(1974) é uma medida de qualidade de ajuste penalizada pela complexidade do modelo (número de parâmetros), por isso em comparação de modelos o objetivo é buscar o AIC de menor valor, selecionando o modelo que minimiza a divergência, do conflito entre a qualidade do ajuste do modelo e a simplicidade do modelo, o AIC é obtido pela expressão:

$$\text{AIC} = -2(\hat{l}) + 2p, \tag{2.12}$$

onde,$(\hat{l})$ é a log-verossimilhança maximizada e p o número de parâmetros estimados.

3 RESULTADOS E DISCUSSÃO

Neste trabalho, presume-se que a variável resposta Y_i segue a distribuição gama $Y_i \sim G(\mu_i, \phi)$, a distribuição binomial negativa $Y_i \sim BN(\mu_i, \phi)$, e a distribuição normal inversa $Y_i \sim NI(\mu_i, \phi)$ para analisar qual distribuição melhor se ajusta aos dados. Inicialmente, o estudo incluiria a comparação com o modelo de distribuição de poisson, entretanto, esse apresentou resultados com valores de AIC altos e discrepantes em relação aos demais. Portanto, optou-se por comparar as três distribuições que demonstraram melhor ajuste aos dados, utilizando a seleção de variáveis pelo método *stepwise* com critério AIC.

Foram realizadas modelagens para as distribuições utilizando diferentes funções de ligação, tais como inversa, logarítmica, identidade e inversa quadrada. Contudo, nos ajustes para as distribuições, as diferentes funções de ligação não demonstraram diferenças significativas nos valores de AIC, mantendo os mesmos coeficientes. Assim, para os modelos apresentados neste trabalho, os preditores lineares referem-se aos modelos gama, binomial negativa e normal inversa com a função de ligação identidade, simplificando as interpretações.

A variável *população* é mencionada neste trabalho apenas com propósitos descritivos, devido à sua alta correlação com a variável de interesse, quantidade de casos notificados por Covid-19. O aumento da população está diretamente associado ao aumento na quantidade de casos. Portanto, essas variáveis são autoexplicativas e, por essa razão, é removida na criação dos modelos para evitar multicolinearidade. Além disso, nenhuma das macrorregioes apresentou dados faltantes ou outliers significativos, que prejudicassem os modelos.

3.1 Macrorregião 1

Na Tabela 4 encontram-se as análises descritivas das variáveis em estudo. Pode-se visualizar que na Macrorregião 1 houve uma média de 3.463 casos notificados de Covid-19 por município, sendo o mínimo com 207 e o máximo com 103.940, esse valor máximo refere-se aos casos da capital João Pessoa, desse modo, pode-se notar que é um *outlier*, um valor discrepante dos demais, sendo bastante distante ainda em comparação do 3º quartil com 2.386 casos por cidade. Na variável independente *pib* tem-se um intervalo amplo entre o valor mínimo de 4,42 e o valor máximo de 29,6 esse fato, deve-se a diversidade de municípios com pequenas e grandes populações presentes na Macrorregião 1, os demais índices não apresentam um intervalo destacante entre os valores mínimos e máximos.

Nos dados da Macrorregião 1, ao realizar o teste de correlação entre variáveis, apresentou o valor de 0,8 de correlação entre a variável *renda* e *educação*, o que pode comprometer o ajuste do modelo aos dados. Para selecionar qual variável remover, foi feito o teste de multicolinearidade pelo VIF, em que a variável *renda* obteve um valor maior que 5, por isso, para os dados da Macrorregião 1, a variável *renda* foi removida para prosseguir os ajustes.

Tabela 4 – Macro 1: Análise descritiva das variáveis explicativas e variável resposta

Variável	Média	Desvio Padrão	Mínimo	Máximo	1º quartil	3º quartil
casos	3.463	13.052	207	103.940	588	2.386
gini	0,507	0,046	0,42	0,7	0,485	0,52
edu	0,462	0,058	0,379	0,693	0,426	0,48
long	0,751	0,034	0,675	0,832	0,729	0,773
pib	7,43	4,55	4,42	29,6	5,23	7,10
rdep	58,2	5,21	41	67,3	55,5	61,5

Fonte: Elaborada pela autora, 2023.

Na Tabela 5, apresentam-se os preditores lineares obtidos para os dados da Macrorregião 1 nas distribuições gama, binomial negativa e normal inversa. Observa-se que o modelo com menor valor de AIC é o gama. Entretanto, ao compararmos a quantidade de variáveis independentes incluídas nos modelos, percebe-se que, para a distribuição normal inversa, o método de seleção de variáveis *stepwise* incluiu uma variável a mais, a *longevidade*. Consequentemente, o valor de AIC será maior. Realizou-se, portanto, uma investigação de comparação entre as distribuições com mais critérios. Ao criar modelos que incluem todas as variáveis independentes, a gama apresentou o menor valor de AIC. Além disso, quando comparado pelo gráfico QQPlot, a gama também apresentou melhor ajuste aos dados. Em vista disso, a gama será a distribuição utilizada no ajuste final para os dados da Macrorregião 1, não havendo problemas de multicolinearidade.

Tabela 5 – Critério de informação AIC pelo método *stepwise* para a distribuição Gama, Binomial Negativa e Normal Inversa ajustados aos dados de casos notificados para Macrorregião 1

Distribuição	Preditor Linear	AIC
Gama	$\eta_i = \beta_0 + \beta_1(gini) + \beta_2(edu) - \beta_3(rdep)$	1032,19
Binomial Negativa	$\eta_i = \beta_0 + \beta_1(gini) + \beta_2(edu) + \beta_3(rdep)$	1032,23
Normal Inversa	$\eta_i = \beta_0 + \beta_1(gini) + \beta_2(edu) - \beta_3(rdep) + \beta_4(long)$	1037,84

Fonte: Elaborada pela autora, 2023.

Na Tabela 6, encontram-se as estimativas dos parâmetros para o modelo final escolhido, na distribuição gama, contendo as variáveis índice de *gini*, *educação* e *razão de dependência*, todas significativas ao nível de 1%.

Na Figura 1, tem-se o gráfico QQplot, gráfico normal de probabilidade para os resíduos, apresentando que os resíduos estão normalmente distribuídos, portanto, confirma-se que não há indícios de afastamento da suposição de distribuição gama com função de ligação identidade para a variável *casos*, portanto, conclui-se que o modelo final para os casos notificados da Macrorregião 1 é dado por:

$$\eta_i = 7,764 + 4,381(gini) + 6,607(edu) - 0,098(rdep) \tag{3.1}$$

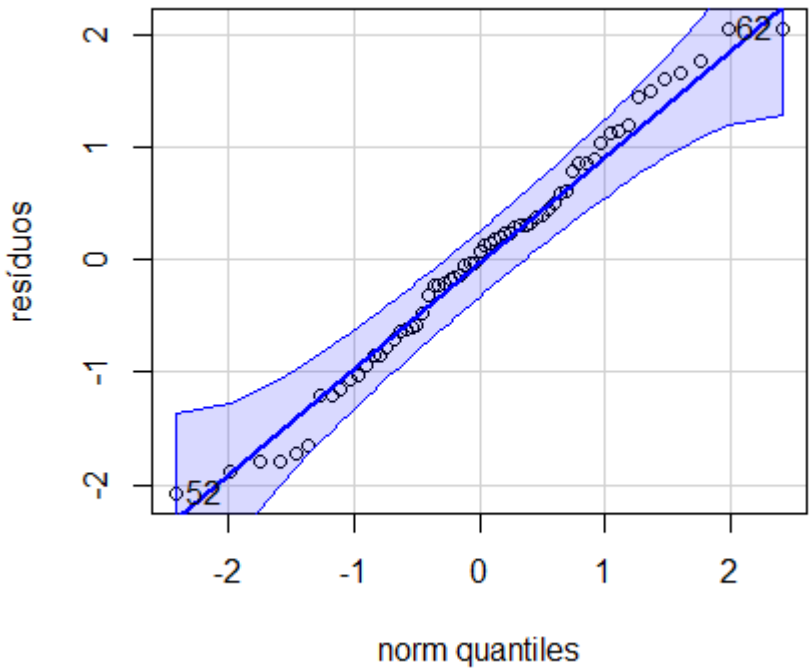
Tabela 6 – Estimativas dos parâmetros para a distribuição gama ajustado aos dados de casos notificados para Macrorregião 1

Coeficientes	Estimativas	Erro padrão	Valor-p
Constante	7,764	2,140	<0,001
gini	4,381	2,024	<0,001
edu	6,607	2,211	<0,001
rdep	-0,098	0,023	<0,001

Fonte: Elaborada pela autora, 2023.

Compreende-se que de acordo com o modelo ajustado aos dados da Macroregião 1, considerando que todas as variáveis independentes sejam iguais a zero, o valor esperado de η_i é de 7,764, assim, para cada aumento em uma unidade do índice de *gini* há um aumento esperado de 4,381 na média da distribuição da quantidade de casos, da mesma forma, para cada aumento em uma unidade do índice de *educação* há um aumento esperado na média da distribuição da quantidade de casos de 6,607. Já para a variável *razão de dependência*, para cada aumento em uma unidade da variável há uma diminuição esperada na média da distribuição da quantidade de casos de 0,098.

Figura 1 – Gráfico QQPlot referente ao modelo ajustado aos casos notificados por Covid-19 na Macrorregião 1



Fonte: Elaborada pela autora, 2023.

3.2 Macrorregião 2

Na Tabela 7 encontram-se as análises descritivas das variáveis em estudo para a Macrorregião 2, pode-se visualizar que houve uma média de 1.637 casos notificados de Covid-19 por cidade, sendo o mínimo com 109 e o máximo com 42.778. Assim como observou-se na macrorregião 1, tem-se que há uma exceção na variável *pib* apresentando um valor mínimo de 3,81 e um valor máximo de 79,6 enquanto, nas demais variáveis observa-se que não há valores destoantes quando comparamos os valores máximos e mínimos.

Tabela 7 – Macro 2: Análise descritiva das variáveis explicativas e variável resposta

Variável	Média	Mediana	Desvio Padrão	Mínimo	Máximo	1º quartil	3º quartil
casos	1.637	640	5.222	109	42.778	331	1.261
gini	0,495	0,49	0,04	0,42	0,58	0,46	0,525
edu	0,482	0,476	0,054	0,373	0,654	0,44	0,514
long	0,754	0,754	0,026	0,699	0,812	0,736	0,77
renda	0,567	0,567	0,034	0,491	0,702	0,545	0,585
pib	5,99	5,68	1,57	4,38	13,6	4,98	6,40
rdep	57,7	57,9	3,89	46,1	65,6	55,1	60,3

Fonte: Elaborada pela autora, 2023.

Na Tabela 8, constata-se que o preditor linear com menor valor de AIC, consequentemente, o melhor ajustado aos dados foi da Normal Inversa, logo, será a distribuição utilizada como modelo final para os dados da Macrorregião 2, não havendo problemas de multicolinearidade.

Tabela 8 – Critério de informação AIC pelo método *stepwise* para a distribuição Gama, Binomial Negativa e Normal Inversa ajustados aos dados de casos notificados para Macrorregião 2

Distribuição	Preditor Linear	AIC
Gama	$\eta_i = \beta_0 + \beta_1(gini) + \beta_2(renda) + \beta_3(long) - \beta_4(edu) + \beta_5(pib)$	1041,67
Binomial Negativa	$\eta_i = \beta_0 + \beta_1(gini) + \beta_2(renda) + \beta_3(long) + \beta_4(edu) + \beta_5(pib)$	1041.74
Normal Inversa	$\eta_i = \beta_0 + \beta_1(gini) + \beta_2(renda) + \beta_3(long) - \beta_4(edu) + \beta_5(rdep)$	1033,18

Fonte: Elaborada pela autora, 2023.

Na Tabela 9, apresenta-se as estimativas dos parâmetros para o modelo final selecionado, incluindo as variáveis índice de *gini, renda, longevidade,educação* e *razão de dependência*. A maioria dessas variáveis mostrou-se significativa a um nível de 1%, exceto *educação*, que possui um valor-p de 0,120, sugerindo que a variável não é significativa para o modelo, porém ela apresentou valor de 0,4 de correlação com a variável resposta, indicando que *edu* tem influência positiva na variação de *casos*.Além disso, um modelo de teste sem *educação* foi criado, resultando em um AIC de 1033,49, com valor superior ao modelo completo. Portanto, a variável *educação* foi mantida no ajuste final.

Tabela 9 – Estimativa dos parâmetros para a distribuição Normal Inversa ajustado aos dados de casos notificados para Macrorregião 2

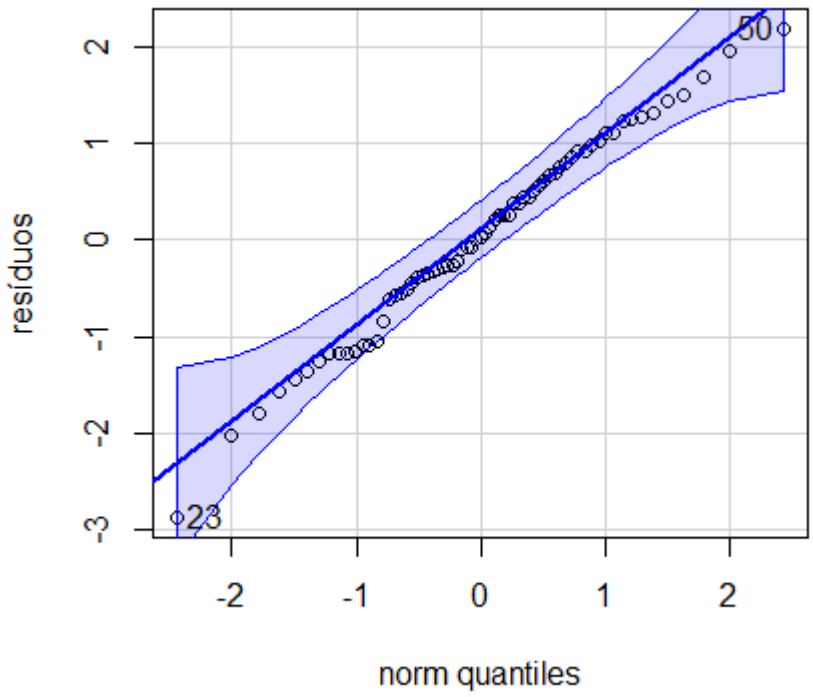
Coeficientes	Estimativas	Erro padrão	Valor-p
constante	-23,305	5,064	<0,001
gini	11,624	2,801	<0,001
renda	14,717	4,854	<0,001
long	14,364	3,531	<0,001
edu	-3,751	2,381	0,120
rdep	0,090	0,039	<0,001

Fonte: Elaborada pela autora, 2023.

Na Figura 2, tem-se o gráfico QQplot, gráfico normal de probabilidade para os resíduos, comprova-se que não há indícios de afastamento da suposição de distribuição normal inversa com função de ligação identidade para a variável casos, portanto conclui-se que o modelo final para os casos notificados da Macrorregião 2 é dado por:

$$\eta_i = -23,305 + 11,624(gini) + 14,717(renda) + 14,364(long) - 3,751(edu) + 0,090(rdep) \tag{3.2}$$

Figura 2 – Gráfico QQPlot referente ao modelo ajustado aos casos notificados por Covid-19 na Macrorregião 2



Fonte: Elaborada pela autora, 2023.

Verifica-se pelo modelo final ajustado que, considerando que todas as variáveis independentes sejam iguais a zero, é esperado um valor de η_i de -23,305, assim, para cada aumento em uma unidade do índice de *gini* há um aumento esperado na média da distribuição da quantidade de casos de 11,624, da mesma forma, para cada aumento em uma unidade do índice de *renda* há um aumento esperado na média da distribuição da quantidade de casos de 14,717, para cada aumento em uma unidade do índice de *longevidade*(saúde) há um aumento esperado na média da distribuição da quantidade de casos de 14,364, enquanto para cada aumento em uma unidade o do índice de *educação* há uma diminuição esperada na média da distribuição da quantidade de casos de 3,751, e para cada aumento em uma unidade do índice de *razão de dependência* há um aumento esperado de 0,090 na média da distribuição da quantidade de casos.

3.3 Macrorregião 3

Na Tabela 10, observa-se as análises descritivas das variáveis em estudo. Pode-se visualizar que na Macrorregião 3 houve uma média de 4.112 casos notificados de Covid-19 por cidade, sendo o mínimo com 60 e o máximo com 13.830. Constata-se que a Macro 3, diferente das Macrorregiões 1 e 2, não apresenta nenhuma variável explicativa com um intervalo de valor amplo quando comparados os valores máximos e mínimos. Entende-se pelo 3º quartil que apenas 25% dos valores de casos notificados são maiores que 885, isso deve-se ao fato que em sua maioria, os municípios que compõem a Macrorregião 3 são de populações pequenas.

Tabela 10 – Macro 3: Análise descritiva das variáveis explicativas e variável resposta

Variável	Média	Mediana	Desvio Padrão	Mínimo	Máximo	1º quartil	3º quartil
casos	4.112	492	2.038	60	13.830	279	885
gini	0,499	0,5	0,043	0,4	0,65	0,46	0,53
edu	0,484	0,478	0,055	0,38	0,714	0,452	0,513
long	0,76	0,762	0,031	0,672	0,821	0,744	0,781
renda	0,567	0,564	0,034	0,488	0,668	,0542	0,583
pib	5,75	5,36	1,30	3,81	11,2	4,92	6,16
rdep	55,1	54,5	4,66	47,5	79,6	52,6	56,8

Fonte: Elaborada pela autora, 2023.

Na Tabela 11, constata-se que o preditor linear com menor valor de AIC, consequentemente, o melhor ajustado aos dados foi da Normal Inversa, logo, será a distribuição utilizada como modelo final para os dados da Macrorregião 3, não havendo problemas de multicolinearidade.

Tabela 11 – Critério de informação AIC pelo método *stepwise* para a distribuição Gama, Binomial Negativa e Normal Inversa ajustados aos dados de casos notificados para Macrorregião 3

Distribuição	Preditor Linear	AIC
Gama	$\eta_i = \beta_0 + \beta_1(gini) + \beta_2(renda) - \beta_4(edu) + \beta_5(long)$	1381,76
Binomial Negativa	$\eta_i = \beta_0 + \beta_1(gini) + \beta_2(renda) + +\beta_3(edu) + \beta_4(pib)$	1381,85
Normal Inversa	$\eta_i = \beta_0 + \beta_1(gini) + \beta_2(renda) - \beta_4(edu) + \beta_5(rdep)$	1381,65

Fonte: Elaborada pela autora, 2023.

Na Tabela 12 encontram-se as estimativas dos parâmetros para o modelo final escolhido, com as variáveis índice de *gini, renda, educação* e *razão de dependência*, com a maioria significativa ao nível de 1%, por exceção da *razão de dependência* que apresenta um valor-p de 0,121 o que poderia indicar que a variável não é significativa para o modelo, por isso foi realizado um estudo focado na relação desta variável com a quantidade de *casos* notificados, em que apresentaram uma correlação de -0,4 demonstrando que a *razão de dependência* contribui para a variação na quantidade de *casos*, também foi criado um modelo de teste retirando a variável *rdep*, obtendo o AIC 1381,814 apontando um valor maior quando comparado ao modelo que inclui a variável, consequentemente, para o ajuste final decide-se manter a variável explicativa.

Tabela 12 – Estimativa dos parâmetros para a distribuição Normal Inversa ajustado aos dados de casos notificados para Macrorregião 3

Coeficientes	Estimativas	Erro padrão	Valor-p
constante	-4,219	2,176	<0,001
gini	8,184	2,872	<0,001
renda	17,510	3,395	<0,001
edu	-3,749	1,780	<0,001
rdep	-0,025	0,016	0,121

Fonte: Elaborada pela autora, 2023.

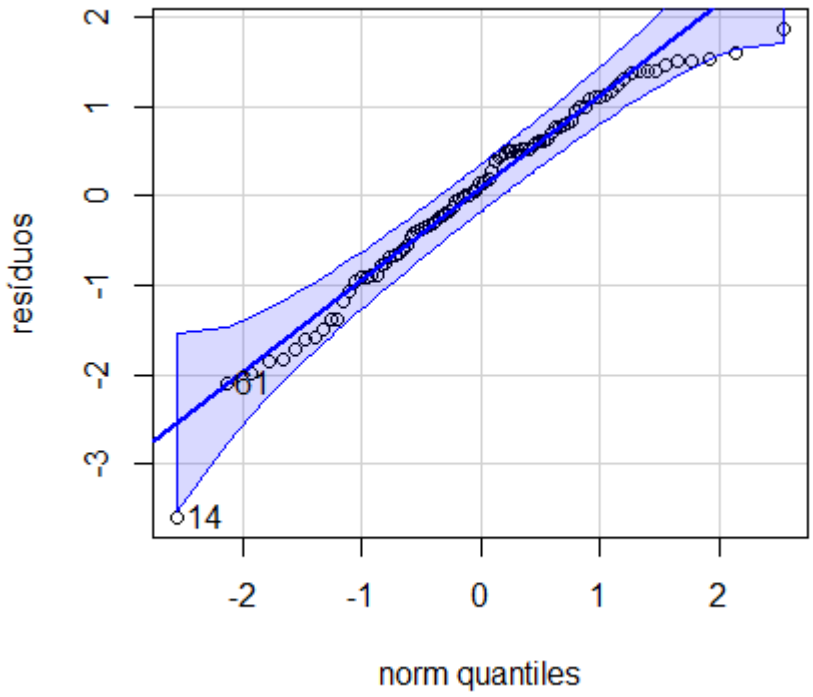
Na Figura 3, observa-se o gráfico QQplot, gráfico normal de probabilidade para os resíduos, comprova-se que não há indícios de afastamento da suposição de distribuição normal inversa com função de ligação identidade para a variável casos, portanto conclui-se que o modelo final para os casos notificados da Macrorregião 3 é dado por:

$$\eta_i = -4,219 + 8,184(gini) + 17,510(renda) - 3,749(edu) - 0,025(rdep) \tag{3.3}$$

Entende-se que, considerando que todas as variáveis independentes sejam iguais a zero, o valor esperado de η_i é -4,219, assim, para cada aumento em uma unidade do índice de *gini* há um aumento esperado na média da distribuição da quantidade de casos de 8,184, da mesma forma, para cada aumento em uma unidade do índice de *renda* há um aumento esperado na média da distribuição da quantidade de casos de 17,510. Já para a variável *educação* tem-se que, para cada

aumento em uma unidade da variável há uma diminuição esperada na média da distribuição da quantidade de casos de 3,749, da mesma forma, para cada aumento em uma unidade da *razão de dependência* há uma diminuição esperada na média da distribuição da quantidade de casos de 0,025.

Figura 3 – Gráfico QQPlot referente ao modelo ajustado aos casos notificados por Covid-19 na Macrorregião 3



Fonte: Elaborada pela autora, 2023.

Na Tabela 13, apresenta-se os três modelos finais para cada Macrorregião do estado da Paraíba. Verifica-se que o índice de *gini* foi significativo para os três modelos, *renda* e *educação* estiveram presentes em dois modelos, a variável *longevidade* foi significativa apenas no ajuste aos dados da Macrorregião 2, o índice *razão de dependência* também foi significativo nos três modelos, porém apenas no modelo 2 que houve uma mudança de sinal, sugerindo que na Macro 2 houve mais diferenças no modelo final quando comparado as outras duas Macrorregiões, isso deve-se ao fato de que é a população mais diversificada entre as três macrorregiões.

Na análise descritiva das três Macrorregiões compreende-se que os índices de *gini*, *educação* e *longevidade* apresentam valores de desvio padrão menores, indicando que os dados tem baixa variação em torno da média, nas variáveis *pib* e *razão de dependência* tem-se maiores valores de desvio padrão, indicando que há uma grande variação nos dados. Para a variável quantidade de *casos* apresenta-se um valor de desvio padrão alto, justificado pela diversidade de populações municipais.

Tabela 13 – Comparação entre os modelos por Macrorregião

Macro	Preditor Linear
1	$\eta_i = 7,764 + 4,381(gini) + 6,607(edu) - 0,098(rdep)$
2	$\eta_i = -23,305 + 11,624(gini) + 14,717(renda) + 14,364(long) + 0,090(rdep)$
3	$\eta_i = -4,219 + 8,184(gini) + 17,510(renda) - 3,749(edu) - 0,025(rdep)$

Fonte: Elaborada pela autora, 2023.

Nesta análise, não obteve-se nenhum modelo final que inclui a variável *pib* como influenciadora na quantidade de casos, apesar de na análise de correlação indicar que há uma correlação positiva entre o PIB e os casos notificados. Para Lippi et al. (2020) observando regiões da Itália, um dos países mais impactados pela pandemia, demonstra uma relação positiva e significativa entre o PIB per capita e o número de óbitos por Covid-19, sendo um indicativo de que o comportamento de indicadores econômicos pode contribuir para entender a dinâmica de evolução desta doença. Para ele, questões como a alta poluição industrial, obesidade e hipertensão (típicos de regiões mais ricas economicamente na Itália) ajudam a explicar essa correlação.

Conforme o estudo de Packer et. al. (2021), entre os 25 países com população mais contaminadas por Covid-19, observou-se que a quantidade de casos e mortes foram mais elevados em nações territorialmente maiores e com maior PIB. Em relação à mortalidade, identificou uma relação inversamente proporcional ao PIB per capita. Os resultados encontrados no artigo sugerem que os determinantes da Covid-19 podem ser explicados por fatores sociais, econômicos, demográficos e comportamentais culturais.

O estudo realizado por Gelfand et. al. (2021), investigou variáveis relacionadas as normas sociais que estariam supostamente associadas ao sucesso dos países na limitação de casos e mortes por Covid-19, utilizando regressão de mínimos quadrados ordinários, em que o índice de Gini foi significativo nos modelos ajustados, corroborando com os resultados obtidos neste estudo, além disto, o artigo destaca a importância de compreender o que explica esta variação de quantidade de casos, em que, não é importante apenas para o avanço da teoria, mas também para orientar intervenções destinadas a enfrentar futuras ameaças coletivas.

Os casos notificados por Covid-19 na Paraíba, seguem uma trajetória semelhante ao estado do Acre, segundo Assis et. al. (2021), revelando que a quantidade de casos tem a concentração predominante na capital, especialmente em bairros de maior IDH, habitados por pessoas de classe média e alta. Os demais casos estão distribuídos na região metropolitana e interior, com uma tendência de disseminação menos concentrada nos municípios do estado. Análogo aos resultados que Borzacchiello e Maria (2020) encontraram no estado do Ceará, e Garcia et. al. (2021) no estado do Paraná. Indicando uma relação direta entre os casos confirmados de Covid-19, sua evolução, o IDH e o número de habitantes, demonstrando um aumento proporcional em cada município.

Para Belchior et.al. (2022) locais com IDHM mais baixo mostram uma quantidade maior de casos confirmados, além disso, o índice apresentou estar inversamente relacionado aos óbitos, indicando que municípios com IDHM mais alto registram menos óbitos em São Paulo, a análise aponta para uma relação inversa entre o índice e a probabilidade de contaminação, revelando que residentes em áreas caracterizadas por condições socioeconômicas precárias e acesso limitado a serviços urbanos, apresentam taxas mais elevadas de incidência de Covid-19.

Amaral e Aguiar (2020) enfatizam que a Covid-19, apesar de ser uma pandemia global, não afeta igualmente todos os territórios e grupos socioeconômicos. Em uma cidade já caracterizada por profundas desigualdades, a pandemia pode agravar ainda mais essas disparidades, destacando a necessidade de ações específicas para proteger as populações mais marginalizadas. Dito isso, análises como esta, associando variáveis socioeconômicas e sociodemográficas da Paraíba são relevantes para aprofundar estudos sobre a variação de casos por Covid-19, de acordo com as macrorregiões de saúde possibilitando a adoção de diferentes políticas públicas para cada realidade populacional. O estudo de Mann et. al. (1992) também já destacou esses indicadores sociais, argumentando que fatores de natureza social e econômica desempenham um papel significativo na propagação de epidemias.

De acordo com Lima e Alves (2020), a renda per capita interfere no aumento das taxas de contaminação por Covid-19, enquanto para o indicador de escolaridade do IDHM tem uma relação inversa com a taxa de contaminação, indicando que distritos com maior acesso à educação, possuem menores índices de confirmação da doença. Já para a longevidade, ainda, conforme Lima e Alves (2020) é demonstrada a ideia de que onde a população é mais saudável, com uma expectativa maior de vida, há menos mortes pela Covid-19, distintivamente do ajuste obtido neste estudo, para a macrorregião 2, que foi o único modelo final incluindo a variável *longevidade*, apresentou coeficiente positivo sugerindo que influencia no aumento de casos.

De acordo com Hoffman et. al. (2016) é fundamental compreender os efeitos que determinadas variáveis exercem, ou aparentam exercer, sobre outras. Mesmo na ausência de uma relação de causa e efeito entre essas variáveis, é possível estabelecer uma relação por meio de uma expressão matemática.

Conforme destacado por Apolinário (2022), é relevante ressaltar a importância dos modelos gama e normal inversa no contexto dos Modelos Lineares Generalizados (MLGs), em contraste com os modelos lineares convencionais, uma vez que estes últimos tem premissas de normalidade dos erros e homogeneidade de variâncias para um bom ajuste aos dados. Assim como Barros (2016) pontuou que, outra vantagem dos MLGs é o fato de não precisar fazer transformações nas variáveis respostas e/ou explicativas, facilitando, em muitos casos, a interpretação dos resultados. Embora, segundo Cordeiro e Demétrio (2008) não é incomum nos MLGs casos em que os dados são primeiramente transformado para depois seguir o ajuste de modelo.

4 CONCLUSÃO

A análise realizada reforça a necessidade de considerar contextos regionais singulares ao criar políticas de saúde pública, reconhecendo que a dinâmica da pandemia é influenciada por diversos fatores socioeconômicos e/ou sociodemográficos. Sendo assim, este estudo evidencia a complexidade das relações entre variáveis independentes e a disseminação de casos de Covid-19 nas diferentes macrorregiões da Paraíba.

Os modelos finais destacam a importância do índice de Gini em todos ajustes, mas outras variáveis, como a renda, a educação e a longevidade, variam em relevância e em como cada índice pode exercer uma influência positiva ou negativa na quantidade de casos notificados.

Desta forma, por meio de modelos estatísticos, utilizando preditores lineares, foi possível obter o melhor modelo ajustado, em que a escolha de modelos gama e normal inversa com função de ligação identidade nos Modelos Lineares Generalizados revela-se apropriada, proporcionando ajustes eficazes sem a necessidade de transformações complexas nas variáveis.

REFERÊNCIAS

AKAIKE, Hirotugu. *A new look at the statistical model identification*. Institute of Statistical Matheniatie, Ainato-ku, Tokyo, Japan. 1974.

ALVES , Marileide F. et. al. *Seleção de variáveis stepwise aplicadas em redes neurais artificiais para previsão de demanda de cargas elétricas*. Departamento de Matemática- UNESP, SP. 2013.

AMARAL, Lucas; AGUIAR, Rafael. *Neighborhood Effects and Urban Inequalities: The Impact of Covid-19 on the Periphery of Salvador, Brazil*. City e Society. 2020.

APOLINÁRIO, Lindembergson. *Uso dos modelos gama e normal inverso para a análise de dados positivos assimétricos*. 2022. 79 f. Trabalho de Conclusão de Curso (Graduação em Estatística) – Universidade Federal do Ceará, Fortaleza. 2022.

ASSIS, Edmilson et.al. *Evolução da COVID-19 da semana epidemiológica 16 a 53 em um Estado da Amazônia Ocidental-Acre/Brasil em 2020*. Cardiorespiratory Research Group, Department of Biological and Medical Sciences, Oxford Brookes University, Headington Campus, OX3 0BP, United Kingdom. 2021

BARROS, Danilo Cordeiro. *Ajuste de modelos lineares generalizados para dados positivos assimétricos*. Trabalho de Conclusão de Curso (Graduação em Estatística), Universidade Estadual da Paraíba, Campina Grande. 2016.

BELCHIOR, et. al. *Análise da relação entre o índice IDHM e a densidade demográfica com a incidência de Covid-19 no município de São Paulo - SP*. Revista Brasileira de Planejamento e Desenvolvimento. 2022.

BORZACCHIELO, José; MARIA, Alexsandra. *Pandemia do Coronavírus no Brasil: Impactos no Território Cearense*. Espaço e Economia, Ceará. 2020

BRANDÃO, Isabel Cristina et. al. *Análise da Organização da Rede de Saúde da Paraíba a Partir do Modelo de Regionalização*. Revista Brasileira de Ciências de Saúde, João Pessoa,v. 16, n.3, p.367-352. 2012

BRENO, Sávio et. al. *Pandemia da COVID-19: o maior desafio do século XXI*. Universidade Federal do Vale do São Francisco (UNIVASF), Paulo Afonso, Bahia. 2020.

CAMPOS, M. R. et. al. *Carga de doença da COVID-19 e de suas complicações agudas e crônicas: reflexões sobre a mensuração (DALY) e perspectivas no Sistema Único de Saúde*. Rio de Janeiro. 2020.

CEFOR. *Plano Estadual de Educação Permanente em Saúde do Estado da Paraíba 2019 – 2022.* João Pessoa. 2019.

CLARICE, et.al. *Análise de modelos de regressão linear com aplicações*. Editora da UNICAMP, Campinas., São Paulo. 1999.

CORDEIRO, G. M.; DEMÉTRIO. *Modelos lineares generalizados e extensões.* C. G. São Paulo. 2008.

DATASUS. *Razão de dependência.* Disponível em: http://tabnet.datasus.gov.br/tabdata/LivroIDB/2edrev/a16.pdf. Acesso em: 20 out. 2023

DOBSON, A. J.; BARNETT. *An introduction to generalized linear models.* A. CRC Press. 2008.

DRAPER, N. R.; SMITH, H. *Applied regression analysis*. John Wiley and Sons, Inc. Canada . 1998

GARCIA,Nathália et. al. *Análise temporo-espacial da evolução da Covid-19 no Estadodo Paranáno período de março a setembro de 2020*. Brazilian Journal of Development, Curitiba, v.7, n.4. 2021

GELFAND, M. J. et. al. *The relationship between cultural tightness–looseness and COVID-19 cases and deaths: a global analysis*. 2021.

GORI, Alexandre Maia. *Econometria: conceitos e aplicações*. *E-book*. 2017.

HOFFMAN, R. et. al. *Análise de regressão: uma introdução à econometria.* São Paulo. 2016.

IBGE. *Produto Interno Bruto - PIB*. Disponível em:https://www.ibge.gov.br/explica/pib. Acesso em: 20 out. 2023

IWATA, Alexandre Xavier; SANDOVAL, Geraldo Góes *Introdução ao software R e análise econométrica. E-book*. 2019.

LACERDA, Fábio Henrique De Souza. *A importância da estatística descritiva na pandemia de Covid-19.*, Revista Científica Multidisciplinar Núcleo do Conhecimento. Ano 05, Ed. 08, Vol. 02, pp. 05-14. 2020.

LIMA , Tiago Pessoa; ALVES, Joao Guilherme. *COVID-19 lethality in non-elderly individuals in cities with different Human Development Index*.Tropical Doctor, v. 51, n. 1, p. 124-125. 2021.

LIPPI, G. et. al. *The death rate for COVID-19 is positively associated with gross domestic products. Acta Bio-Medica*. Atenei Parmensis. 2020.

LUCAMBIO, Pérez Fernando. *Teoria estatística para modelos lineares generalizados*. Universidade Federal do Paraná. 2022.

MINISTÉRIO DE SAÚDE. *Guia de vigilância epidemiológica : emergência de saúde pública de importância nacional pela doença pelo coronavírus 2019 –covid-19*. Secretaria de Vigilância em Saúde. Brasília: Ministério da Saúde, 2022.

MALVEIRA, Emilly. *Modelos lineares generalizados: Análise de Dados Categóricos*. Universidade Federal de Minas Gerais. 2018.

MANN, Jonathan M. et. al. *AIDS in the World.* Harvard University Press. 1992.

MYERS, R.H.; MONTGOMERY, D.C. *Response surface methodology: Process and Product Optimization Using Designed Experiments. 2 Ed., John Wiley Sons*. New York, NY. 2002

NELDER JA, Wedderburn RWM. *Generalized linear models.* J R Stat Soc A . 1972.

PACKER, Dyenifer et. al. *Influência das características dos países na disseminação da Covid-19*. Revista Gestão Organizacional, Chapecó. 2021.

PAULA, Gilberto. *Modelos de regressão com apoio computacional*. Universidade de São Paulo. 2013.

RODRIGUES, S. C. A. *Modelo de regressão linear e suas aplicações.* Tese (Doutorado), Universidade da Beira Interior. 2012.

SOUSA, Rafaela. *Índice de Desenvolvimento Humano (IDH)*. Brasil escola. Disponível em: https://brasilescola.uol.com.br/geografia/idh-indice-desenvolvimento-humano.htm. Acesso em 20 out. 2023.

SCHMIDT, C. M. C. *Modelo de regressão de Poisson aplicado à área da saúde.* Mestrado em Modelagem Matemática, Universidade Regional do Noroeste do Estado do Rio Grande do Sul, Ijúi. 2003.

TURKMAN, LOIOLA. *Modeleos lineares generalizados: Da teoria à práticas*. Universidade de Lisboa. 2000.

WOLFFENBUTTEL, Andréa. *O que é Índice de Gini?*. IPEA. 2004. Disponível em: https://www.ipea.gov.br/desafios/index.php. Acesso em: 20 out. 2023