UNIVERSIDADE FEDERAL DO AMAZONAS

INSTITUTO DE CIÊNCIAS EXATAS

DEPARTAMENTO DE ESTATÍSTICA

BACHARELADO EM ESTATÍSTICA

ABRAÃO AUDILLE DE SOUZA LIMA

MODELO DE REGRESSÃO BETA APLICADO À INCIDÊNCIA DA COVID-19 NO BRASIL

MANAUS – AMAZONAS

2023

ABRAÃO AUDILLE DE SOUZA LIMA

MODELO DE REGRESSÃO BETA APLICADO À INCIDÊNCIA DA COVID-19 NO BRASIL

Monografia apresentada ao curso de graduação em Estatística da Universidade Federal do Amazonas, como requisito parcial à obtenção do grau de bacharel em Estatística.

Orientador: Prof. Dr. Jhonnata Bezerra de Carvalho

MANAUS – AMAZONAS

2023

Ficha Catalográfica

Ficha catalográfica elaborada automaticamente de acordo com os dados fornecidos pelo(a) autor(a).

Lima, Abraão Audille de Souza
L732m Modelo de regressão beta aplicado à incidência da covid-19 no Brasil / Abraão Audille de Souza Lima . 2023
40 f.: il. color; 31 cm.

Orientador: Jhonnata Bezerra de Carvalho
TCC de Graduação (Estatística) - Universidade Federal do Amazonas.

1. Bootstrap. 2. Logito. 3. Máxima Verossimilhança. 4. Probito. I. Carvalho, Jhonnata Bezerra de. II. Universidade Federal do Amazonas III. Título

ABRAÃO AUDILLE DE SOUZA LIMA

MODELO DE REGRESSÃO BETA APLICADO À INCIDÊNCIA DA COVID-19 NO BRASIL

Monografia apresentada ao curso de graduação em Estatística da Universidade Federal do Amazonas, como requisito parcial à obtenção do grau de bacharel em Estatística.

Aprovada em: 02 de março de 2023

BANCA EXAMINADORA

Documento assinado digitalmente
gov.br
JHONNATA BEZERRA DE CARVALHO
Data: 18/05/2023 17:08:05-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Jhonnata Bezerra de Carvalho (Orientador)
Instituto de Ciências Exatas - ICE
Universidade Federal do Amazonas - UFAM

Documento assinado digitalmente
gov.br
JOSE MIR JUSTINO DA COSTA
Data: 17/05/2023 20:28:20-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. José Mir Justino da Costa
Instituto de Ciências Exatas - ICE
Universidade Federal do Amazonas - UFAM

Documento assinado digitalmente
gov.br
THEMIS DA COSTA ABENSUR LEAO
Data: 18/05/2023 11:32:02-0300
Verifique em https://validar.iti.gov.br

Profª Ma. Themis da Costa Abensur Leão
Instituto de Ciências Exatas - ICE
Universidade Federal do Amazonas - UFAM

AGRADECIMENTOS

Agradeço a Deus por ter me dado a oportunidade de estar concluindo esse curso e ter me sustentado em todos os momentos. Ao meu orientador, professor Dr. Jhonnata Bezerra, pela paciência que teve comigo, pelos conselhos, dedicação e esforço para me ajudar a concluir este trabalho. Aos professores do departamento de Estatística da UFAM que foram essenciais na transmissão do conhecimento. Agradeço aos meus pais que sempre investiram em mim e me incentivaram nos estudos. A minha noiva pelas palavras de incentivo, motivação e foi um suporte em todos os momentos na escrita deste trabalho. E a todas as pessoas que contribuíram de alguma forma para a minha formação acadêmica.

RESUMO

O modelo de regressão beta, em muitas situações, é utilizado para modelar dados que apresentam o suporte definido no intervalo $(0, 1)$. Este modelo ganhou bastante destaque desde a sua apresentação por Ferrari e Cribari-Neto (2004). Geralmente, é utilizado o método da máxima verossimilhança (MV) para estimar os parâmetros do modelo beta, entretanto, quando o tamanho da amostra é pequeno, os vieses dos estimadores não podem ser desprezados, sendo necessária a correção dos vieses. Dentre as correções existentes para os vieses dos estimadores de MV, pode-se citar a correção via *bootstrap*. Neste trabalho, um estudo de simulação foi realizado para verificar o desempenho dos estimadores de MV e de MV corrigidos via *bootstrap* dos parâmetros do modelo de regressão beta. Além disso, foi realizado uma aplicação do modelo de regressão beta aos dados de COVID-19 dos 26 estados brasileiros, incluindo o Distrito Federal, no ano de 2020.

Palavras-chave: *Bootstrap*. Logito. Máxima Verossimilhança. Probito.

ABSTRACT

The beta regression model is often used to model data that has support defined on the interval $(0, 1)$. This model gained significant attention since its presentation by Ferrari e Cribari-Neto (2004). Maximum likelihood (ML) method is generally used to estimate the parameters of the beta regression model, however, when the sample size is small, the biases of the estimators cannot be neglected. Therefore, it is necessary to correct these biases. Among the existing corrections for ML estimators, the bootstrap correction can be cited. In this text, simulation study was conducted to verify the performance of ML estimators and bootstrap-corrected ML estimators of the parameters of the beta regression model. Additionally, the beta regression model was applied to COVID-19 data from the 26 brazilians states, including the Federal District, in 2020.

Keywords: Bootstrap. Logit. Maximum Likelihood. Probit.

LISTA DE TABELAS

SUMÁRIO

LISTA DE FIGURAS

1 INTRODUÇÃO

Modelos de regressão são técnicas estatísticas que investigam e modelam o relacionamento entre variáveis. Francis Galton (1822-1911) foi o pioneiro nos estudos dessa técnica ao analisar a relação entre a altura de pais e filhos. Galton descobriu que pais altos, tenderiam a ter filhos mais baixos e pais mais baixos tenderiam a ter filhos mais altos e essa relação ficou conhecida como regressão à média. Os estudos acerca dessa técnica evoluíram ao longo dos anos e com isso trouxe diversas aplicações nas ciências biológicas, física e química, ciências sociais, engenharias e ciências econômicas.

No modelo de regressão linear tradicional, que assume-se ter distribuição normal é comum realizar transformações na variável resposta e/ou nas variáveis explicativas, o que pode acarretar em uma série de dificuldades para interpretar os parâmetros do modelo. Houve então o surgimento de uma classe de modelos de regressão que podem contornar esse problema de falta de interpretabilidade. Nesse contexto, o modelo de regressão beta foi proposto por Ferrari e Cribari-Neto (2004) em que os autores reparametrizam a distribuição beta em termos da média e de um parâmetro de precisão. A aplicação desse modelo é a mais indicada para dados que representam taxas e proporções, ou seja, restritos ao intervalo (0,1). Ospina, Cribari-Neto e Vasconcellos (2006) propuseram correções para os estimadores de máxima verossimilhança e aplicaram o modelo em dados de petróleo.

Segundo Ferrari e Cribari-Neto (2004), embora seja possível transformar uma variável resposta que pertença ao intervalo (0,1) utilizando o modelo de regressão linear normal, essa transformação apresenta limitações, pois ao realizar a transformação ocorre a dificuldade de interpretar o modelo em termo da variável resposta original.

Uma situação muito frequente na prática é encontrar dados com um tamanho amostral pequeno. Nesse caso, pode acontecer que o estimador de máxima verossimilhança seja viesado, que é a situação dos estimadores dos parâmetros da distribuição beta reparametrizada. Para contornar este problema é indicado o uso de métodos que corrigem o viés e um dos métodos que podem ser utilizados é o método *bootstrap*, fazendo com que esse estimador apresente menor viés em comparação com o estimador de máxima verossimilhança. Além disso, é possível fazer uma análise numérica desses estimadores obtendo resultados importantes, como: coeficiente de assimetria, curtose, erro-padrão, média, variância, erro quadrático médio, dentre outros.

A análise de diagnóstico do modelo de regressão beta reparametrizada apresentada posteriormente inclui a análise gráfica dos resíduos, pois através desta análise é possível verificar

se os resíduos estão bem ajustados ao modelo. Ademais, uma medida importante que é utilizada para averiguar a qualidade do ajuste do modelo é o pseudo R^2, denotado por R^2_p, proposto por Ferrari e Cribari-Neto (2004), que verifica o quão próximo estão os dados da linha de regressão ajustada. Além disso, faz-se necessário utilizar os critérios de seleção AIC (Critério de Informação de Aike), que foi proposto por Akaike (1973) e o BIC (Critério de Informação Bayesiano) proposto inicialmente por Schwarz (1978), ambos os métodos são uma alternativa para escolha de um modelo apropriado que explique melhor os dados.

Dentro deste cenário um exemplo de aplicação é utilizado para verificar a influência de fatores socioeconômicos, demográficos, epidemiológicos e da estrutura do sistema de saúde na evolução da pandemia da COVID-19 no Brasil, tendo como período desde o primeiro caso em 26 de fevereiro de 2020 até o dia 23 de agosto de 2020.

1.1 OBJETIVOS

1.1.1 Objetivo Geral

Estudar os principais aspectos do modelo de regressão beta com dispersão variável.

1.1.2 Objetivos específicos

- Apresentar conceitos básicos da distribuição beta;
- Realizar um estudo de simulação para verificar o desempenho dos estimadores para os parâmetros do modelo;
- Utilizar uma correção de viés para os estimadores de MV;
- Avaliar o desempenho dos estimadores com correção de viés aos dados de COVID-19 no Brasil em 2020.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 DISTRIBUIÇÃO BETA

A distribuição beta é uma distribuição de probabilidade contínua que possui o suporte definido no intervalo (0,1). Essa distribuição é muito utilizada para dados que apresentam taxas e proporções. Júnior *et al.* (1995) utilizaram essa distribuição para analisar o conjunto de dados de velocidade média dos ventos da região de Botucatu em São Paulo de uma série de 20 anos, o objetivo era verificar se a ocorrência desses eventos meteorológicos seriam adequadamente interpretados pela distribuição beta.

A função de densidade de probabilidade de uma variável aleatória Y, com distribuição beta e parâmetros de forma $\alpha > 0$ e $\beta > 0$, é dada por

$$f(y;\alpha,\beta) = \frac{y^{\alpha-1}(1-y)^{\beta-1}}{B(\alpha,\beta)}, \qquad 0 < y < 1, \tag{2.1}$$

em que $B(\alpha,\beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta)$ é a função beta e $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1}e^{-t}dt$ é a função gama. Utilizaremos a seguinte notação $Y \sim beta(\alpha,\beta)$ para nos referirmos à distribuição beta.

Segundo Silva (2020) a distribuição beta possui algumas características importantes, por exemplo quando α e β são iguais, a distribuição apresenta simetria; quando α é menor que β, ocorre assimetria à direita; e a assimetria à esquerda acontece quando α é maior que β.

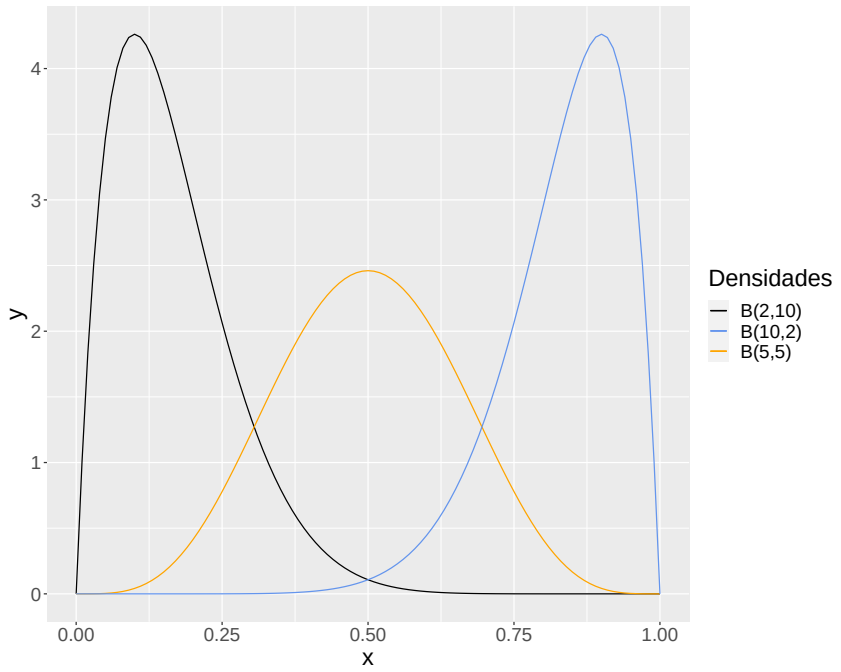

Figura 1 – Densidades de probabilidade da distribuição beta

A Figura 1, exibe os gráficos das densidades da distribuição beta com diferentes características, como: assimetria à direita, simetria e assimetria à esquerda.

A função de distribuição de Y é dada por

$$F(y;\alpha,\beta) = \frac{B(y;\alpha,\beta)}{B(\alpha,\beta)}, \tag{2.2}$$

em que $B(y;\alpha,\beta) = \int_0^y t^{\alpha-1}(1-t)^{\beta-1}dt$ é a função beta incompleta. A esperança e variância da distribuição beta são dadas por

$$E(Y) = \frac{\alpha}{\alpha+\beta} \quad e \quad Var(Y) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

2.2 DISTRIBUIÇÃO BETA REPARAMETRIZADA

A reparametrização da distribuição beta proposta por Ferrari e Cribari-Neto (2004) tem como objetivo modelar a média da variável aleatória, para isso é preciso fazer manipulações tanto na média quanto na variância da distribuição beta. Seja $\mu = \alpha/(\alpha+\beta)$ e $\phi = \alpha+\beta$, segue que

$$\frac{\alpha}{\phi} = \mu \Rightarrow \alpha = \mu\phi$$

e, de forma semelhante, temos que para o parâmetro β

$$\phi = \mu\phi + \beta \Rightarrow \beta = (1-\mu)\phi.$$

Ao substituir α e β em $E(Y) = \alpha/(\alpha+\beta)$, temos que $E(Y) = \mu$. Além disso, para a variância temos que

$$\begin{aligned} Var(Y) &= \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = \frac{\mu\phi(1-\mu)\phi}{(\mu\phi+(1-\mu)\phi)^2(\mu\phi+(1-\mu)\phi+1)} \\ &= \frac{\phi^2\mu(1-\mu)}{\phi^2(\phi+1)} = \frac{\mu(1-\mu)}{\phi+1} = \frac{V(\mu)}{1+\phi}, \end{aligned}$$

em que $V(\mu)$ é a função de variância. Ao aplicar os valores de α e β na equação (2.1), tem-se que a função densidade reparametrizada é dada por

$$f(y;\mu,\phi) = \frac{y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}}{B(\mu\phi,(1-\mu)\phi)}, \qquad 0<y<1, \tag{2.3}$$

no qual $B(\mu\phi,(1-\mu)\phi) = \Gamma(\mu\phi)\Gamma((1-\mu)\phi)/\Gamma(\mu\phi+(1-\mu)\phi)$ é a função beta reparametrizada, $\Gamma(\mu\phi) = \int_0^\infty t^{\mu\phi-1}e^{-t}dt$ é a função gama reparametrizada, $0<\mu<1$ e $\phi>0$.

2.3 MODELO DE REGRESSÃO BETA

Sejam as $Y_1, \ldots, Y_n$ variáveis aleatórias independentes, em que cada Y_i, $i = 1, \ldots, n$, possui a densidade da equação (2.3), com a média μ_i e o parâmetro de precisão ϕ_i. Assim, o modelo de regressão beta com dispersão variável pode ser escrito como

$$g(\mu_i) = \mathbf{x}_i^\top \boldsymbol{\beta} = \eta_i \quad \text{e} \quad h(\phi_i) = \mathbf{z}_i^\top \boldsymbol{\gamma} = \vartheta_i, \tag{2.4}$$

em que $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ e $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_q)^\top$ são os vetores de coeficientes desconhecidos da regressão, no qual $\boldsymbol{\beta} \in \mathbb{R}^p$ e $\boldsymbol{\gamma} \in \mathbb{R}^q$ com $p + q < n$, η_i e ϑ_i são os preditores lineares, $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\top$ e $\mathbf{z}_i = (z_{i1}, \ldots, z_{iq})^\top$ são os vetores de observações conhecidos de dimensões p e q, respectivamente, para $i = 1, \ldots, n$. Além disso, as funções de ligação $g_1 : \mathbb{R} \to \mathbb{R}^+$ e $g_2 : \mathbb{R} \to \mathbb{R}^+$, na equação (2.4), são funções estritamente monótonas, positivas e pelo menos duas vezes diferenciáveis. Existem várias possibilidades de escolha para g e h, por exemplo, as funções: logito, probito e complemento log-log para g; e as funções: logaritmo, raiz quadrada e inversa para h.

2.3.1 Estimação dos parâmetros

Segundo Ospina (2007), utiliza-se a estimação de máxima verossimilhança para encontrar os estimadores da regressão beta com dispersão variável. Com isso, segue que o logaritmo da função de verossimilhança é dado por

$$\ell(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \sum_{i=1}^{n} \ell_i(\mu_i, \phi_i),$$

com

$$\begin{aligned} \ell_i(\mu_i, \phi_i) &= \log\Gamma(\phi_i) - \log\Gamma(\mu_i\phi_i) - \log\Gamma((1-\mu_i)\phi_i) + (\mu_i\phi_i - 1)\log y_i \\ &+ \{(1-\mu_i)\phi_i - 1\}\log(1-y_i). \end{aligned} \tag{2.5}$$

Logo, para $t = 1, \ldots, k$, temos que a função escore, que corresponde à derivada do o logaritmo da função de verossimilhança em relacão a β_t, é dada por

$$\frac{\partial \ell(\boldsymbol{\beta}, \boldsymbol{\gamma})}{\partial \beta_t} = \sum_{i=1}^{n} \frac{\partial \ell_i(\mu_i, \phi_i)}{\partial \mu_i} \frac{d\mu_i}{d\eta_i} \frac{\partial \eta_i}{\partial \beta_t}, \tag{2.6}$$

com $d\mu_i/d\eta_i = 1/g'(\mu_i)$ e

$$\frac{\partial \ell_i(\mu_i,\phi_i)}{\partial \mu_i} = \phi_i \left[\log\left(\frac{y_i}{1-y_i}\right) - \{\psi(\mu_i\phi_i) - \psi((1-\mu_i)\phi_i)\}\right]. \tag{2.7}$$

Definindo o termo

$$\mu_t^* = \psi(\mu_i\phi_i) - \psi((1-\mu_i)\phi_i), \tag{2.8}$$

segue que

$$\frac{\partial \ell(\boldsymbol{\beta},\boldsymbol{\gamma})}{\partial \beta_t} = \sum_{i=1}^{n} \phi_i(y_i^* - \mu_i^*)\frac{1}{g'(\mu_i)}x_{it}, \tag{2.9}$$

com y_i^* sendo o logito de y_i. O vetor escore é representado por

$$U_\beta(\boldsymbol{\beta},\boldsymbol{\gamma}) = \mathbf{X}^\top \boldsymbol{\Phi}\mathbf{T}(\mathbf{y}^* - \boldsymbol{\mu}^*)$$

em que $\boldsymbol{\Phi} = diag\{\phi_1,\ldots,\phi_n\}$ é uma matriz diagonal com valores de ϕ_i e $\boldsymbol{\mu}^* = (\mu_1^*,\ldots,\mu_n^*)^\top$ é o vetor de valores μ_i^*.

Segundo Ospina (2007), ao considerar as derivadas do logaritmo da função verossimilhança em relação aos parâmetros de γ_j, $j = 1,\ldots,q$, obtem-se

$$\frac{\partial \ell(\boldsymbol{\beta},\boldsymbol{\gamma})}{\partial \gamma_j} = \sum_{i=1}^{n} \frac{\partial \ell_i(\mu_i,\phi_i)}{\partial \phi_i}\frac{d\phi_i}{d\vartheta_i}\frac{\partial \vartheta_i}{\partial \gamma_j}, \tag{2.10}$$

com $d\phi_i/d\vartheta_i = 1/h'(\phi_i)$. Logo,

$$\begin{aligned}\frac{\partial \ell_i(\mu_i,\phi_i)}{\partial \phi_i} &= \mu_i\left[\log\frac{y_i}{1-y_i} - (\psi(\mu_i\phi_i) - \psi((1-\mu_i)\phi_i))\right] + \\ &\quad \log(1-y_i) - \psi((1-\mu_i)\phi_i) + \psi(\phi_i).\end{aligned} \tag{2.11}$$

Com isso, a função escore de cada um dos parâmetros γ_j é representada por

$$\frac{\partial \ell(\boldsymbol{\beta},\boldsymbol{\gamma})}{\partial \gamma_j} = \sum_{i=1}^{n} \left[\mu_i(y_i^* - \mu_i^*) + \log(1-y_i) - \psi((1-\mu_i)\phi_i) + \psi(\phi_i)\right]\frac{1}{h'(\phi_i)}z_{ij},$$

que pode ser reescrita como

$$\frac{\partial \ell(\boldsymbol{\beta},\boldsymbol{\gamma})}{\partial \gamma_j} = \sum_{i=1}^{n} a_i\frac{1}{h'(\phi_i)}z_{ij},$$

em que

$$a_i = \mu_i(y_i^* - \mu_i^*) + \log(1-y_i) - \psi((1-\mu_i)\phi_i) + \psi(\phi_i). \tag{2.12}$$

Com isso, o vetor escore pode ser escrito da forma

$$\boldsymbol{U}_{\boldsymbol{\gamma}}(\boldsymbol{\beta},\boldsymbol{\gamma}) = \mathbf{Z}^\top \mathbf{H}\mathbf{a},$$

em que $\mathbf{Z}$ é uma matriz $n \times q$ cuja i-ésima linha apresenta as observações $\boldsymbol{z}_i^\top$, além disso

$$\mathbf{H} = diag\left\{\frac{1}{h'(\phi_1)}, \ldots, \frac{1}{h'(\phi_n)}\right\} \tag{2.13}$$

e $\mathbf{a} = (a_1, \ldots, a_n)^\top$. Segundo Ospina (2007) é possível encontrar a matriz de informação de Fisher conjunta dos vetores de $\boldsymbol{\beta}$ e $\boldsymbol{\gamma}$ ao calcular a derivada segunda do logaritmo da função de verossimilhança, sendo assim para $t = 1, \ldots, p$, temos que

$$\frac{\partial^2 \ell(\boldsymbol{\beta},\boldsymbol{\gamma})}{\partial \beta_t \partial \beta_r} = \sum_{i=1}^{n} \frac{\partial}{\partial \mu_i}\left(\frac{\partial \ell_i(\mu_i,\phi_i)}{\partial \mu_i}\frac{d\mu_i}{d\eta_i}\right)\frac{d\mu_i}{d\eta_i}\frac{\partial \eta_i}{\partial \beta_j}x_{it}.$$

Uma vez que $E(\partial \ell_i(\mu_i,\phi_i)/\partial \mu_i) = 0$,

$$E\left(\frac{\partial^2 \ell(\boldsymbol{\beta},\boldsymbol{\gamma})}{\partial \beta_t \partial \beta_r}\right) = \sum_{i=1}^{n} E\left(\frac{\partial^2 \ell_i(\mu_i,\phi_i)}{\partial \mu_i^2}\right)\left(\frac{d\mu_i}{d\eta_i}\right)^2 x_{it}x_{ir}. \tag{2.14}$$

Com isso, segue que a equação (2.14) pode ser escrita como

$$E\left(\frac{\partial^2 \ell(\boldsymbol{\beta},\boldsymbol{\gamma})}{\partial \beta_t \partial \beta_r}\right) = -\sum_{i=1}^{n} \phi_i \omega_i x_{it} x_{ir}, \tag{2.15}$$

em que ω_i é o peso associado e definido por

$$\omega_i = \phi_i\left\{\psi'(\mu_i\phi_i) + \psi'((1-\mu_i)\phi_i)\right\}\frac{1}{\{g'(\mu_i)\}^2}. \tag{2.16}$$

Portanto a equação (2.15) pode ser escrita na forma matricial

$$E\left(\frac{\partial^2 \ell(\boldsymbol{\beta},\boldsymbol{\gamma})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top}\right) = -\boldsymbol{X}^\top \boldsymbol{\Phi} \boldsymbol{W} \boldsymbol{X},$$

em que $\mathbf{W} = diag\{\omega_1, \ldots, \omega_n\}$. Sendo assim, segue da equação (2.9) que as derivadas segundas de $\ell(\boldsymbol{\beta},\boldsymbol{\gamma})$ em relação a β_t e γ_j, com $t = 1, \ldots, k$ e $j = 1, \ldots, q$ são dadas por

$$\frac{\partial^2 \ell(\boldsymbol{\beta},\boldsymbol{\gamma})}{\partial \beta_t \partial \gamma_j} = \sum_{i=1}^{n} \frac{\partial}{\partial \phi_i}\left(\phi_i(y_i^* - \mu_i^*)\frac{1}{g'(\mu_i)x_{it}}\right)\frac{d\phi_i}{d\vartheta_i}\frac{\partial \vartheta_i}{\partial \gamma_j},$$

logo,

$$\begin{aligned}\frac{\partial^2 \ell(\boldsymbol{\beta},\boldsymbol{\gamma})}{\partial \beta_t \partial \gamma_j} &= \sum_{i=1}^{n} \left\{ (y_i^* - \mu_i^*) - \phi_i \left[\psi'(\mu_i\phi_i)\mu_i - \psi'((1-\mu_i)\phi_i)(1-\mu_i) \right] \right\} \\ &\times \frac{1}{g'(\mu_i)} \frac{1}{h'(\phi_i)} x_{it} z_{ij}.\end{aligned}$$

Uma vez que $E(y_i^* - \mu_i^*) = 0$, temos que

$$E\left(\frac{\partial^2 \ell(\boldsymbol{\beta},\boldsymbol{\gamma})}{\partial \beta_t \partial \gamma_j} \right) = -\sum_{i=1}^{n} c_i \frac{1}{g'(\mu_i)} \frac{1}{h'(\phi_i)} x_{it} z_{ij}. \tag{2.17}$$

em que $c_i = \phi_i [\psi'(\mu_i\phi_i)\mu_i - \psi'((1-\mu_i)\phi_i)(1-\mu_i)]$.

Além disso, segundo Ospina (2007), a equação (2.17) pode ser escrita na forma matricial, como sendo

$$E\left(\frac{\partial^2 \ell(\boldsymbol{\beta},\boldsymbol{\gamma})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\gamma}} \right) = -\mathbf{X}^\top \mathbf{CTHZ},$$

com $\mathbf{C} = diag\{c_1, \ldots, c_n\}$.

A partir da equação (2.10) é possível encontrar as derivadas segundas de $\ell(\boldsymbol{\beta},\boldsymbol{\gamma})$ em relação a γ_j e γ_l com $j,l = 1,\ldots,q$, com isso segue que

$$\frac{\partial^2 \ell(\boldsymbol{\beta},\boldsymbol{\gamma})}{\partial \gamma_j \partial \gamma_l} = \sum_{i=1}^{n} \left(\frac{\partial^2 \ell_i(\mu_i,\phi_i)}{\partial \phi_i^2} \frac{d\phi_i}{d\vartheta_i} + \frac{\ell_i(\mu_i,\phi_i)}{\partial \phi_i} \frac{\partial}{\partial \phi_i} \left(\frac{d\phi_i}{d\vartheta_i} \right) \right) \frac{d\phi_i}{d\vartheta_i} z_{il} z_{ij}.$$

Uma vez que $E(\partial \ell_i(\mu_i,\phi_i)/\partial \phi_i) = 0$, então

$$E\left(\frac{\partial^2 \ell(\boldsymbol{\beta},\boldsymbol{\gamma})}{\partial \gamma_j \partial \gamma_l} \right) = \sum_{i=1}^{n} E\left(\frac{\partial^2 \ell_i(\mu_i,\phi_i)}{\partial \phi_i^2} \right) \left(\frac{d\phi_i}{d\vartheta_i} \right)^2 z_{il} z_{ij}.$$

A partir da equação (2.11), encontramos

$$E\left(\frac{\partial^2 \ell_i(\mu_i,\phi_i)}{\partial \phi_i^2} \right) = -\left[\psi'(\mu_i\phi_i)\mu_i^2 + \psi'((1-\mu_i)\phi_i)(1-\mu_i)^2 - \psi'(\phi_i) \right].$$

Com isso, temos que

$$\begin{aligned}E\left(\frac{\partial^2 \ell(\boldsymbol{\beta},\boldsymbol{\gamma})}{\partial \gamma_j \partial \gamma_l} \right) &= -\sum_{i=1} \left[\psi'(\mu_i\phi_i)\mu_i^2 + \psi'((1-\mu_i)\phi_i)(1-\mu_i)^2 - \psi'(\phi_i) \right] \\ &\times \frac{1}{\{h'(\phi_i)\}^2} z_{il} z_{ij},\end{aligned}$$

assim, Ospina (2007) demostra que essa equação pode ser escrita na forma matricial, tal que

$$E\left(\frac{\partial^2 \ell(\boldsymbol{\beta},\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}^\top} \right) = -\boldsymbol{Z}^\top \boldsymbol{D}^* \boldsymbol{Z},$$

em que $\boldsymbol{D}^* = diag\{d_i^*, \ldots, d_n^*\}$, com

$$d_i^* = \left[\psi'(\mu_i\phi_i)\mu_i^2 + \psi'((1-\mu_i)\phi_i)(1-\mu_i)^2 - \psi'(\phi_i)\right] \frac{1}{\{h'(\phi_i)\}^2}. \tag{2.18}$$

Portanto, segue que a matriz informação de fisher é dada por

$$\boldsymbol{K}^* = \boldsymbol{K}^*(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \begin{pmatrix} \boldsymbol{K}_{\boldsymbol{\beta}\boldsymbol{\beta}}^* & \boldsymbol{K}_{\boldsymbol{\beta}\boldsymbol{\gamma}}^* \\ \boldsymbol{K}_{\boldsymbol{\gamma}\boldsymbol{\beta}}^* & \boldsymbol{K}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}^* \end{pmatrix},$$

em que $\boldsymbol{K}_{\boldsymbol{\beta}\boldsymbol{\beta}}^* = \boldsymbol{X}^\top \boldsymbol{\Phi} \boldsymbol{W} \boldsymbol{X}$, $\boldsymbol{K}_{\boldsymbol{\beta}\boldsymbol{\gamma}}^* = \boldsymbol{K}_{\boldsymbol{\gamma}\boldsymbol{\beta}}^{*\top} = \boldsymbol{X}^\top \boldsymbol{C} \boldsymbol{T} \boldsymbol{H} \boldsymbol{Z}$ e $\boldsymbol{K}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}^* = \boldsymbol{Z}^\top \boldsymbol{D}^* \boldsymbol{Z}$.

Sob certas condições de regularidade, para tamanho de amostras grandes, a distribuição conjunta de $\hat{\boldsymbol{\beta}}$ e $\hat{\boldsymbol{\gamma}}$ é aproximadamente normal (k+q) multivariada, é dada por

$$\begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\gamma}} \end{pmatrix} \sim N_{k+q}\left(\begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\gamma} \end{pmatrix}, \boldsymbol{K}^{*-1}\right) \tag{2.19}$$

em que $\hat{\boldsymbol{\beta}}$ e $\hat{\boldsymbol{\gamma}}$ são os estimadores de máxima verossimilhança, respectivamente.

De acordo com Pawitan (2001), pode-se utilizar o resultado da equação (2.19) para realizar testes de hipóteses, isto é, seja $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\gamma})^\top$ um vetor de parâmetros, então temos que $H_0 : \theta_i = \theta_i^0$ *versus* $H_1 : \theta_i \neq \theta_i^0$, em que θ_i^0 é um parâmetro especificado para o parâmetro desconhecido de θ_i, para $i = 1, 2, \ldots, k+q$. A estatística de teste do teste de hipótese é dada por $Z = \left(\hat{\theta}_i - \theta_i^0\right) / \sqrt{k^{ii}}$, em que k^{ii} é i-ésimo elemento da diagonal principal de $\boldsymbol{K}^{*-1}$. É importante ressaltar que sob H_0 e para n grande, temos que $\mathbf{Z}$ possui, aproximadamente uma distribuição normal padrão.

Um resultado importante para a equação (2.19) é a construção do intervalo de confiança para o parâmetro θ_i^0. Aproximando $(1-\alpha)100\%$ o intervalo de confiança para θ_i^0 é definido como sendo $\left(\hat{\theta}_i^0 - z_{1-\alpha/2}\sqrt{k^{ii}}; \hat{\theta}_i^0 + z_{1-\alpha/2}\sqrt{k^{ii}}\right)$, em que $\Phi(z_{1-\alpha/2}) = 1 - \alpha/2$.

2.3.2 Correção de viés pelo método bootstrap

Os métodos *bootstrap* foram introduzidos por Efron (1979) e são considerados uma classe de métodos de Monte Carlo, cujo objetivo é estimar a distribuição de uma população por reamostragem. Este método é bastante útil quando se pretende avaliar, por exemplo, para um estimador, o seu viés, o seu erro padrão, além disso quando se pretende estimar a distribuição de probabilidade deste estimador.

2.3.3 Método de Monte Carlo

Antes de aprofundarmos no método *bootstrap* é preciso primeiro entender como são definidas as simulações de Monte Carlo. De acordo com Hammersley e Handscomb (1964) o nome ”Monte Carlo” foi introduzido para indicar um método que foi utilizado no projeto Manhattan, responsável por desenvolver as bombas atômicas na Segunda Guerra Mundial, cujos pioneiros na utilização desses métodos foram Stanislaw Ulam, von Neumann e Fermi. O método parte do princípio da convergência de probabilidade, de acordo com (CASELLA; BERGER, 2021) a covergência de probabilidade é definida como: Sejam $Y_1, Y_2, \ldots$ uma sequência de variáveis aleatórias, que convergem para uma variável aleatória Y. Dizemos que Y_n converge em probabilidade para Y, se para cada $\varepsilon > 0$, tivermos

$$\lim_{n\to+\infty} P(\mid Y_n - Y \mid \geq \varepsilon) = 0 \quad ou \quad \lim_{n\to+\infty} P(\mid Y_n - Y \mid < \varepsilon) = 1.$$

Um resultado importante da convergência de probabilidade é a Lei Fraca dos Grandes Números, pois demonstra que a média amostral converge em probabilidade para a média populacional μ. (CASELLA; BERGER, 2021) definem que dada a sequência de variáveis aleatórias $Y_1, Y_2, \ldots$ independentes e identicamente distribuídas, com $E(Y_i) = \mu$ e $Var(Y_i) = 0 < \sigma^2$. Considere $\bar{Y}_n = (1/n)\sum_{i=1}^{n} Y_i$. Então temos que

$$\lim_{n\to+\infty} P(\mid \bar{Y}_n - \mu \mid < \varepsilon) = 1,$$

para todo $\varepsilon > 0$, logo $\bar{Y}_n$ converge em probabilidade para μ. Isso significa que quando n for grande, $\bar{Y}_n$ vai convergir para μ. Com isso, o método de Monte Carlo consiste em gerar amostras $Y = Y_1, Y_2, \ldots, Y_n$, com n grande, a partir dessas amostras, calcula-se a média $\bar{Y}_n$ de forma que $\bar{Y}_n$ converge para μ.

2.3.4 Método Bootstrap

Segundo Efron e Tibshirani (1994) o método *bootstrap* parte do princípio de distribuição empírica, em que y_i, $i = 1, 2, \ldots, n$ são os valores observados, com a probabilidade de $1/n$ para cada y_i, todos obtidos da distribuição empírica F, além disso deseja-se estimar um parâmetro de interesse θ, com base nos dados de y. Uma estimativa para θ é $\hat{\theta} = s(\boldsymbol{y})$.

Com isso o método *bootstrap* pode ser definido como sendo uma amostra aleatória de tamanho n, extraída de uma população com distribuição $\hat{F}$, em que $\mathbf{y}^* = (y_1^*, y_2^*, \ldots, y_n^*)$ é o conjunto de dados *bootstrap*. Outra maneira de definir o método *bootstrap* é afirmar que y^* é uma amostra aleatória de tamanho n, retirada com reposição de uma subpopulação $(y_1, y_2, \ldots, y_n)$ de interesse. O conjunto de dados de y^* corresponde as replicações de *bootstrap* que é denominado $\hat{\theta}^*$, dado por

$$\hat{\theta}_{boot} = s(y^*) \tag{2.20}$$

Nesse caso uma estimativa importante para a equação (2.20) é a média que é obtida por $\bar{y}^* = \sum_{i=1}^{n} y_i^*/n$. Abaixo encontra-se o algoritmo de *bootstrap* que é definido como:

1. Seja uma amostra aleatória observada $\mathbf{y} = (y_1, \ldots, y_n)$ calcule a estatística $\hat{\theta}$ de interesse.
2. A partir da amostra original extraia, com reposição, uma amostra $\mathbf{y}^{*b} = (y_1^{*b}, \ldots, y_n^{*b})$.
3. Calcule a mesma estatística de interesse considerando a amostra bootstrap y^b para obter $\hat{\theta}^{*b}$.
4. Repita as etapas (2) e (3) um número B muito grande de vezes, obtendo $\hat{\theta}^{*1}, \ldots, \hat{\theta}^{*B}$.
5. Use a estimativa da distribuição dada por $\hat{\theta}^{*b}$, para $b = 1, \ldots, B$, para obter as estatísticas desejadas, como: média, erro-padrão, intervalo de confiança, etc.

Acontece que muitas vezes podemos obter um estimador assintoticamente não viesado e isso ocorre principalmente quando o tamanho da amostra de uma população de interesse é significativamente grande, às vezes quando a amostra é pequena o estimador acaba sendo viesado. Neste caso o método *bootstrap* é mais adequado para corrigir o viés, tornando-se uma ferramenta útil para a minimização desse problema.

(MORETTIN; BUSSAB, 2017) o cálculo do viés de um estimador pode ser obtido como: considere $\hat{\theta}$ um estimador de θ, então o cálculo do viés é dado como sendo $b(\hat{\theta}) = E(\hat{\theta}) - \theta$, portanto, dizemos que o estimador é viesado quando $b(\hat{\theta}) \neq 0$ e não viesado quando $b(\hat{\theta}) = 0$.

De forma geral para corrigir o $b(\hat{\theta})$ é preciso estimar o viés, pois $E(\hat{\theta})$ e θ são desconhecidos, com isso, segue que uma estimativa para θ é o próprio $\hat{\theta}$. Outro resultado importante é encontrar a estimativa para $E(\hat{\theta})$, neste caso uma solução plausível é gerar B amostras *bootstrap*, calcular as estimativas $\hat{\theta}$, com $b = 1, \ldots, B$, além disso calcular a média

$$\hat{\theta}_{boot} = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}^{*b},$$

logo, pela Lei Forte dos Grandes Números temos que a média amostral converge probabilisticamente para a média populacional, sendo assim $\hat{\theta}_{boot}$ é uma estimativa para $E(\hat{\theta})$, com isso segue que a estimativa de $b(\hat{\theta})$ é dada por

$$\hat{b}(\hat{\theta}) = \hat{E}(\hat{\theta}) - \hat{\theta},$$

sendo assim o estimador corrigido via *bootstrap* pode ser obtido da seguinte forma

$$\hat{\theta}_{boot} = \hat{\theta} - \hat{V}(\hat{\theta}) = \hat{\theta} - \left[\hat{E}(\hat{\theta}) - \hat{\theta}\right] = 2\hat{\theta} - \hat{E}(\hat{\theta}).$$

Os passos para a correção do viés pelo método *bootstrap* são definidos da seguinte forma:

1. Seja uma amostra aleatória observada $\mathbf{y} = (y_1, \ldots, y_n)$ calcule a estatística $\hat{\theta}$ de interesse.
2. A partir da amostra original extraia, com reposição, uma amostra $\mathbf{y}^{*b} = (y_1^{*b}, \ldots, y_n^{*b})$.
3. Calcule a mesma estatística de interesse considerando a amostra bootstrap $\boldsymbol{y}^b$ para obter $\hat{\theta}^{*b}$.
4. Repita as etapas (2) e (3) um número B muito grande de vezes, obtendo $\hat{\theta}^{*1}, \ldots, \hat{\theta}^{*B}$.
5. Obtenha a estimativa do viés que é dada por $\hat{b}(\hat{\theta}) = \left[\frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^{*b}\right] - \hat{\theta} = \hat{E}(\hat{\theta}) - \hat{\theta}$.
6. Com a estimativa do viés no passo anterior obtenha o estimador corrigido que é dado por $\hat{\theta}_{boot} = \hat{\theta} - \hat{V}(\hat{\theta}) = 2\hat{\theta} - \hat{E}(\hat{\theta})$.

2.3.5 Análise de diagnóstico

Na subseção 2.3 apresentamos o ajuste do modelo de regressão beta com dispersão variável. Ao ajustar esse modelo é preciso analisar a qualidade do ajuste e esse é objetivo desta subseção. Inicialmente, abordaremos a análise de diagnóstico investigando os resíduos quantílicos, em seguida, será utilizado o pseudo R-quadrado ajustado e por último será verificada a seleção do modelo pelo métodos de AIC (Critério de Informação de Aike) e BIC (Critério de Informação Bayesiano).

2.3.6 Resíduos quantílicos

Resíduos, e especialmente gráficos de resíduos, desempenham um papel central na verificação do ajuste de um modelo de regressão. Após o ajuste do modelo é imprescindível a análise residual, pois com isso podemos verificar se os dados apresentam linearidade, a normalidade dos erros, heterocedasticidade, a apresentação de outliers, entre outros, e isso é importante para verificar se o modelo está bem ajustado aos dados. Dentre as diversas análises

residuais veremos neste momento os resíduos quantílicos que foram proposto por Dunn e Smyth (1996).

Segundo Dunn e Smyth (1996) os resíduos quantílicos são definidos como: considere $F(x,\mu,\phi)$ uma distribuição acumulada de uma variável aleatória X. Se F é contínua, então $F(x_i,\mu_i,\phi)$ para $i = 1,2,\ldots,n$, são uniformemente distribuídos no intervalo unitário. Neste caso os resíduos quantílicos são definidos por

$$r_i = \Phi^{-1}(\hat{F}(x_i;\hat{\mu}_i,\hat{\phi}_i))$$

em que $\Phi()$ é a função acumulada da normal padrão, esses resíduos possuem, assintoticamente, distribuição normal padrão.

2.3.7 R-quadrado ajustado

O R-quadrado, ou R^2, é uma medida estatística de quão próximos os dados estão da linha de regressão ajustada. Ele também é conhecido como o coeficiente de determinação ou o coeficiente de determinação múltipla para a regressão múltipla. Segundo Ferrari e Cribari-Neto (2004) em um modelo de regressão beta com dispersão variável o pseudo R^2 ou R^2_p é definido como o quadrado do coeficiente de correlação da amostra entre $\hat{\eta}$ e $g(y)$ e é utilizado como sendo uma medida global de variação explicada, que pode variar entre $0 \leq R^2_p \leq 1$ de forma que quanto mais próximo o R^2_p estiver de 1, significa que os dados estão mais próximo da reta de regressão e que o modelo se ajusta bem aos dados quanto mais próximo R^2_p estiver de 0, significa que existe pouca relação linear e que o modelo não se ajusta bem aos dados.

2.3.8 Critério de seleção de modelos AIC e BIC

A escolha de um modelo apropriado, que explique melhor os dados, é muito importante na análise dos dados. O objetivo principal da seleção de modelos é encontrar um modelo que envolva o mínimo de parâmetros possíveis a serem estimados e que explique bem o comportamento da variável resposta. Nesta subseção será abordado os dois métodos mais utilizados que é o AIC e o BIC.

Proposto por Akaike (1973) o método AIC tem como objetivo estimar a perda de informação, segundo este método, o melhor modelo será aquele com menor valor do AIC, pois vai maximizar o logaritmo da função de verossimilhança e minimizar a perda de informação. Segundo Bayer e Cribari-Neto (2017) o AIC pode ser definido como:

$$AIC = -2\ell(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) + 2k,$$

em que $\hat{\boldsymbol{\beta}}$ e $\hat{\boldsymbol{\gamma}}$ são estimadores de máxima verossimilhança de $\boldsymbol{\beta}$ e $\boldsymbol{\gamma}$, respectivamente. O BIC ou SIC foi introduzido por Schwarz (1978) e é um dos métodos mais populares quando se pretende fazer a seleção de modelos, o BIC tem como critério avaliar a estatística que maximiza a probabilidade de se identificar o verdadeiro modelo entre os avaliados. Segundo Bayer e Cribari-Neto (2017) o BIC é definido como:

$$BIC = -2\ell(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) + k\log(n),$$

com n sendo o número de observações. Segundo esse método é considerado o melhor ajuste o modelo que apresenta o menor BIC.

2.4 ANÁLISE NUMÉRICA

O objetivo desta seção é abordar a análise numérica do estimadores, como verificar: a média, variância, desvio-padrão, erro quadrático médio, viés relativo, assimetria e curtose dos estimadores de MV e MV corrigido via *bootstrap*. O estimador de Monte Carlo é definido como:

$$\hat{\theta}_{mc} = \frac{1}{R}\sum_{k=1}^{R}\hat{\theta}_k$$

em que $\hat{\theta}_k$ é a estimativa de máxima verossimilhança da k-ésima réplica de Monte Carlo e $\hat{\theta}_{mc}$ é a estimativa de Monte Carlo (média das R estimativas), assim podemos calcular o viés do estimador $\hat{\theta}_{mc}$, então temos que

$$b(\hat{\theta}_{mc}) = \hat{\theta}_{mc} - \theta,$$

em que θ é o valor do parâmetro verdadeiro. Uma medida interessante a ser calculada para a avaliação de um estimador pontual é o viés relativo (VR). Ela nos dá uma mensuração percentual de viés, independendo da magnitude do parâmetro. Com isso o viés relativo é definido como

$$VR(\hat{\theta}_{mc}) = \frac{b(\hat{\theta}_{mc})}{\theta} * 100$$

Além do viés, podemos calcular o desvio-padrão do estimador $\hat{\theta}_{mc}$

$$dp(\hat{\theta}_{mc}) = \sqrt{\frac{1}{R-1}\sum_{k=1}^{R}(\hat{\theta}_k - \hat{\theta}_{mc})}.$$

Outra estatística importante para a análise numérica de um estimador é o erro quadrático médio e nesse caso é definido como

$$EQM(\hat{\theta}_{mc}) = \left[dp(\hat{\theta}_{mc})\right]^2 + \left[b(\hat{\theta}_{mc})\right]^2.$$

Uma medida muito importante em se obter é a curtose, pois através dessa medida é possível verificar o quanto os dados são robustos nas extremidades, sendo assim a definição de curtose pode ser empregada no estimador $\hat{\theta}_{mc}$, com isso temos que

$$K = \frac{\frac{\sum_{i=1}^{R}(\theta_i - \bar{\theta}_i)}{R}}{(S^2)^2}, \quad com \quad S^2 = \frac{\sum_{i=1}^{R}(\theta_i - \bar{\theta}_i)}{R-1},$$

com $\bar{\theta}_i$ sendo a média do estimador $\hat{\theta}_{mc}$.

O Coeficiente de Assimetria ou CA é uma medida estatística essencial, pois através dessa estatística é que podemos verificar se os dados estão simétricos em torno do 0, assim pela definição de assimetria temos que

$$CA = \sum_{i=1}^{n}\frac{Z_i^3}{n}, \quad com \quad Z_i = \frac{\theta_i - \bar{\theta}_i}{S}, i = 1,2,\ldots,n.$$

O CA pode ser interpretado como:

1. CA > 0, então a distribuição apresenta assimetria positiva;
2. CA < 0, então a distribuição apresenta assimetria negativa;
3. CA = 0, então a distribuição apresenta simetria perfeita.

É importante salientar que todas essas estatísticas foram calculadas para o estimador $\hat{\theta}_{mc}$, o mesmo vale para o estimador $\hat{\theta}_{boot}$, para isso basta substituir $\hat{\theta}_{mc}$ por $\hat{\theta}_{boot}$.

3 ESTUDO DE SIMULAÇÃO

O estudo de simulação é importante para verificar as definições discutidas nas seções anteriores e para isso foram realizadas, no modelo de regressão beta com dispersão variável definida na equação (2.4), simulações via Monte Carlo com 5000 réplicas, sendo que cada réplica apresenta 1000 réplicas de *bootstrap*, para amostras de tamanho n igual a 20, 27, 30, 40 e 50. Os parâmetros β_0, β_1 e β_1, foram fixados em 1.5, -1 e -1, respectivamente. Os valores de ϕ foram fixados em 5, 15 e 30.

As Tabelas 1, 2 e 3 apresentam as simulações descritas anteriormente, com os valores de $\hat{\theta}_{mc}$ e $\hat{\theta}_{boot}$ sendo o estimador de Monte Carlo e o estimador corrigido do parâmetro θ, respectivamente, com isso foram calculados as estimativas de viés relativo, coeficientes de assimetria e curtose e o desvio-padrão para cada um dos estimadores. Nota-se que em relação as estimativas do parâmetro ϕ é possível verificar que apresentam uma grande variabilidade e aproximando-se poucas vezes do valor original. O contrário ocorre com para as estimativas do parâmetro $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_1)^\top$, neste caso grande parte das estimativas se aproximam de $\boldsymbol{\beta}$. Cabe ressaltar que na maior parte das simulações, as estimativas do estimador $\hat{\theta}_{boot}$ do parâmetro de $\boldsymbol{\beta}$ são as que mais se aproximam do valor verdadeiro.

Em relação aos vieses VR($\hat{\theta}_{mc}$) e VR($\hat{\theta}_{boot}$) que demonstram o percentual do viés, verifica-se que para o parâmetro de $\boldsymbol{\beta}$ o VR($\hat{\theta}_{boot}$) é menor em comparação com o VR($\hat{\theta}_{mc}$), neste caso $\hat{\theta}_{boot}$ é o que está melhor ajustado. Tratando-se do parâmetro ϕ é possível notar que o VR($\hat{\theta}_{boot}$) na maioria dos cenários apresenta menor viés em relação ao VR($\hat{\theta}_{mc}$), a explicação para essa grande diferença deve-se ao fato de $\hat{\theta}_{boot}$ se aproxima mais do parâmetro verdadeiro. Outro fator interessante a destacar é que em várias simulações a medida que o tamanho de n cresce o VR($\hat{\theta}_{boot}$) diminui.

Um fator interessante a ser verificado são os desvios-padrão dp($\hat{\theta}_{mc}$) e dp($\hat{\theta}_{boot}$) para os parâmetros $\boldsymbol{\beta}$ e ϕ, nota-se que em todos os cenários dp($\hat{\theta}_{boot}$) é o que apresenta menor variabilidade, indicando mais homogeneidade. Em relação ao CA($\hat{\theta}_{mc}$) e CA($\hat{\theta}_{boot}$) verifica-se que existe uma variabilidade entre os coeficientes de assimetria dos parâmetros de $\boldsymbol{\beta}$ e ϕ, embora os valores estejam próximo de 0. Outra observação ocorre nas curtoses, que definem o achatamento da curva de probabilidade, na maioria dos casos, temos que as curtoses K($\hat{\theta}_{mc}$) e K($\hat{\theta}_{boot}$) para o parâmetro de $\boldsymbol{\beta}$ estão próximas do valor de referência que é onúmero 3 (curtose da normal), enquanto que para o parâmetro ϕ a distribuição é leptocúrtica.

Tabela 1 – Análise dos estimadores $\hat{\theta}_{mc}$ e $\hat{\theta}_{boot}$ para amostras de tamanho 20 e 27 com β_0, β_1 e β_2 iguais a 1,5, -1, -1 e ϕ iguais a 5, 15 e 30.

n	Parâmetro	θ	$\hat{\theta}_{mc}$	$\hat{\theta}_{boot}$	$VR(\hat{\theta}_{mc})$	$VR(\hat{\theta}_{boot})$	$dp(\hat{\theta}_{mc})$	$dp(\hat{\theta}_{boot})$	$CA(\hat{\theta}_{mc})$	$CA(\hat{\theta}_{boot})$	$K(\hat{\theta}_{mc})$	$K(\hat{\theta}_{boot})$
20	β_0	1,5	1,5510	1,5128	3,4002	-0,8553	0,4813	0,4702	0,1126	0,1219	3,0830	3,0983
	β_1	-1,0	-1,0285	-1,0033	2,8457	-0,3289	0,6843	0,6689	-0,0520	-0,0571	2,9990	3,0083
	β_2	-1,0	-1,0422	-1,0158	4,2212	-1,5789	0,6933	0,6780	-0,0826	-0,0868	3,0240	3,0320
	ϕ	5,0	6,7185	4,4948	34,3695	10,1030	2,5616	1,7038	1,8424	1,8259	10,2039	9,9859
	β_0	1,5	1,5218	1,5070	1,4566	-0,4696	0,2929	0,2899	0,0651	0,0675	3,0493	3,0540
	β_1	-1,0	-1,0177	-1,0076	1,7684	-0,7616	0,4168	0,4130	-0,0160	-0,0174	3,1565	3,1715
	β_2	-1,0	-1,0180	-1,0079	1,8010	-0,7940	0,4255	0,4217	-0,0146	-0,0167	2,9801	2,9733
	ϕ	15,0	19,9315	13,3190	32,8768	11,2067	7,7920	5,2151	1,8340	1,8348	10,0164	9,9732
	β_0	1,5	1,5065	1,4991	0,4352	0,0573	0,2112	0,2100	0,0527	0,0536	2,9819	2,9795
	β_1	-1,0	-0,9994	-0,9945	-0,0551	0,5507	0,2994	0,2981	-0,0126	-0,0145	3,1032	3,1061
	β_2	-1,0	-1,0099	-1,0050	0,9932	-0,4964	0,3038	0,3026	-0,0658	-0,0685	3,1188	3,1209
	ϕ	30,0	40,1148	26,7731	33,7161	10,7562	15,2244	10,1927	1,6079	1,6371	7,7552	7,9850
27	β_0	1,5	1,5297	1,5048	1,9822	-0,3204	0,4945	0,4863	0,1068	0,1123	3,0358	3,0417
	β_1	-1,0	-1,0143	-0,9983	1,4262	0,1703	0,5922	0,5822	-0,0187	-0,0205	2,9721	2,9734
	β_2	-1,0	-1,0275	-1,0111	2,7543	-1,1076	0,5316	0,5228	-0,0029	-0,0080	3,2024	3,1928
	ϕ	5,0	6,0990	4,7378	21,9806	5,2438	1,6851	1,3026	1,0738	1,0684	4,9325	4,9378
	β_0	1,5	1,5039	1,4935	0,2623	0,4353	0,2996	0,2977	0,0501	0,0474	3,2312	3,2309
	β_1	-1,0	-0,9988	-0,9919	-0,1216	0,8076	0,3817	0,3793	0,0061	0,0057	3,1176	3,1232
	β_2	-1,0	-1,0036	-0,9966	0,3624	0,3437	0,3628	0,3603	-0,0480	-0,0488	3,1591	3,1599
	ϕ	15,0	18,2412	14,1227	21,6081	5,8490	5,5537	4,3017	1,3241	1,3359	5,9938	6,0774
	β_0	1,5	1,5070	1,5014	0,4641	-0,0926	0,2170	0,2162	0,1050	0,1061	2,9575	2,9599
	β_1	-1,0	-1,0028	-0,9990	0,2780	0,0981	0,2687	0,2680	-0,0752	-0,0761	2,9846	2,9833
	β_2	-1,0	-1,0053	-1,0015	0,5302	-0,1492	0,2230	0,2222	-0,0912	-0,0916	3,0042	3,0040
	ϕ	30,0	37,1568	28,7335	23,8560	4,2216	11,9290	9,2161	1,6524	1,6441	9,2360	9,2097

Tabela 2 – Análise dos estimadores $\hat{\theta}_{mc}$ e $\hat{\theta}_{boot}$ para amostras de tamanho 30 e 40 com β_0, β_1 e β_2 iguais a 1,5, -1, -1 e ϕ iguais a 5, 15 e 30.

n	Parâmetro	θ	$\hat{\theta}_{mc}$	$\hat{\theta}_{boot}$	$VR(\hat{\theta}_{mc})$	$VR(\hat{\theta}_{boot})$	$dp(\hat{\theta}_{mc})$	$dp(\hat{\theta}_{boot})$	$CA(\hat{\theta}_{mc})$	$CA(\hat{\theta}_{boot})$	$K(\hat{\theta}_{mc})$	$K(\hat{\theta}_{boot})$
30	β_0	1,5	1,5179	1,5033	1,1911	-0,2184	0,3342	0,3295	0,1195	0,0526	4,5075	4,4899
	β_1	-1,0	-1,0113	-1,0016	1,1349	-0,1571	0,4166	0,4110	-0,0607	-0,0258	4,5382	4,5252
	β_2	-1,0	-1,0083	-0,9989	0,8301	0,1067	0,3766	0,3713	-0,0317	-0,0008	4,4138	4,4015
	ϕ	5,0	18,1116	15,8204	262,2321	-216,4089	14,3945	13,0343	0,5381	0,5305	1,8232	1,7898
	β_0	1,5	1,5059	1,4991	0,3904	0,0600	0,2207	0,2195	0,0328	0,0072	3,8765	3,8715
	β_1	-1,0	-1,0027	-0,9983	0,2744	0,1731	0,2742	0,2728	0,0142	0,0250	3,7133	3,7076
	β_2	-1,0	-1,0020	-0,9974	0,1972	0,2614	0,2552	0,2540	-0,0439	-0,0316	3,7092	3,6964
	ϕ	15,0	24,7503	21,1323	65,0021	-40,8820	9,6885	9,1336	0,6238	0,6207	2,8610	2,6647
	β_0	1,5	1,5148	1,5100	0,9874	-0,6693	0,1962	0,1954	-0,0769	-0,0765	3,4467	3,4478
	β_1	-1,0	-1,0033	-1,0002	0,3329	-0,0210	0,2328	0,2319	0,0250	0,0195	3,2456	3,2471
	β_2	-1,0	-1,0142	-1,0110	1,4206	-1,0966	0,2133	0,2124	-0,1049	-0,1024	2,9716	2,9520
	ϕ	30	36,3990	29,1584	21,3301	2,8053	10,3226	8,2631	1,1150	1,1329	5,5210	5,6470
40	β_0	1,5	1,5165	1,4983	1,1030	0,1131	0,3595	0,3551	0,1503	0,1506	2,9544	2,9582
	β_1	-1,0	-0,9964	-0,9846	-0,3635	1,5411	0,4642	0,4593	-0,0098	-0,0062	2,9812	2,9817
	β_2	-1,0	-1,0218	-1,0100	2,1801	-1,0003	0,4401	0,4347	-0,0922	-0,0920	2,9625	2,9632
	ϕ	5,0	5,6412	4,8489	12,8245	3,0213	1,2999	1,1150	0,8723	0,8747	4,1426	4,1471
	β_0	1,5	1,5083	1,5011	0,5557	-0,0723	0,2316	0,2305	0,1442	0,1451	3,2057	3,2098
	β_1	-1,0	-1,0102	-1,0055	1,0214	-0,5489	0,2855	0,2843	-0,0600	-0,0609	2,9462	2,9574
	β_2	-1,0	-1,0034	-0,9985	0,3364	0,1497	0,2873	0,2860	-0,0145	-0,0117	3,2016	3,2090
	ϕ	15,0	17,0690	14,6524	13,7932	2,3172	4,0494	3,4790	1,0422	1,0516	5,0586	5,1236
	β_0	1,5	1,5071	1,5035	0,4720	-0,2355	0,1555	0,1553	0,0550	0,0555	3,0342	3,0371
	β_1	-1,0	-1,0039	-1,0015	0,3892	-0,1500	0,1991	0,1988	0,0102	0,0109	2,9413	2,9375
	β_2	-1,0	-1,0048	-1,0024	0,4819	-0,2407	0,1911	0,1908	-0,0277	-0,0274	2,9871	2,9909
	ϕ	30,0	33,8909	29,5348	12,9697	1,5506	7,7184	6,6917	1,0522	1,0064	5,5572	5,3425

Tabela 3 – Análise dos estimadores $\hat{\theta}_{mc}$ e $\hat{\theta}_{boot}$ para amostras de tamanho 50 com β_0, β_1 e β_2 iguais a 1,5, -1, -1 e ϕ iguais a 5, 15 e 30.

n	Parâmetro	θ	$\hat{\theta}_{mc}$	$\hat{\theta}_{boot}$	$VR(\hat{\theta}_{mc})$	$VR(\hat{\theta}_{boot})$	$dp(\hat{\theta}_{mc})$	$dp(\hat{\theta}_{boot})$	$CA(\hat{\theta}_{mc})$	$CA(\hat{\theta}_{boot})$	$K(\hat{\theta}_{mc})$	$K(\hat{\theta}_{boot})$
50	β_0	1,5	1,5078	1,4935	0,5228	0,4312	0,3345	0,3313	0,0664	0,0676	3,2238	3,2246
	β_1	-1,0	-1,0019	-0,9925	0,1899	0,7505	0,4014	0,3977	-0,0360	-0,0369	3,2747	3,2730
	β_2	-1,0	-1,0083	-0,9988	0,8250	0,1154	0,4214	0,4175	-0,0036	0,0002	3,0808	3,0825
	ϕ	5,0	5,5436	4,9383	10,8722	1,2338	1,1160	0,9918	0,9450	0,9371	5,2055	5,1286
	β_0	1,5	1,5003	1,4945	0,0173	0,3653	0,2000	0,1991	0,0696	0,0765	2,9007	2,9072
	β_1	-1,0	-1,0013	-0,9976	0,1343	0,2434	0,2434	0,2422	-0,0431	-0,0371	3,2311	3,2278
	β_2	-1,0	-0,9973	-0,9937	-0,2680	0,6297	0,2538	0,2526	0,0192	0,0184	2,8671	2,8775
	ϕ	15,0	16,6675	14,8285	11,1168	1,1433	3,6043	3,2061	1,0951	1,1001	5,8163	5,8392
	β_0	1,5	1,5038	1,5007	0,2566	-0,0486	0,1459	0,1456	0,0620	0,0657	3,0275	3,0264
	β_1	-1,0	-1,0013	-0,9992	0,1306	0,0840	0,1890	0,1887	-0,0247	-0,0269	2,9796	2,9714
	β_2	-1,0	-1,0018	-0,9998	0,1840	0,0209	0,1865	0,1860	-0,0307	-0,0322	3,0363	3,0407
	ϕ	30,0	33,2302	29,5526	10,7672	1,4913	6,8797	6,1206	0,7917	0,7855	4,1432	4,1144

3.1 APLICAÇÃO

Um exemplo de aplicação são os dados com 22 variáveis e 27 observações sobre a COVID-19 no Brasil, tendo como período desde o primeiro caso em 26 de fevereiro de 2020 até o dia 23 de agosto de 2020. Além disso, as 27 observações são os 26 Estados brasileiros mais o Distrito Federal. A variável dependente ou variável resposta usada para esta análise será a incidência, esta variável diz respeito a frequência de casos novos de COVID-19 em um intervalo de tempo. Os dados obtidos foram extraídos a partir do artigo escrito por Figueiredo *et al.* (2020).

Uma primeira abordagem dos dados faz-se necessário realizar a análise descritiva da variável resposta, neste caso a Tabela 4 apresenta as principais estatísticas descritivas da variável incidência. Nota-se que a média de incidência está entorno de 0,023963, que equivale a 2.396,3 novos casos por 100 mil habitantes no período de referência. Além disso a maior incidência foi de 6.888,9 mil casos novos por 100 mil habitantes, enquanto que a menor incidência foi de 919,3 casos novos por 100 mil habitantes no período de referência. Além disso, temos que a mediana da incidência é de 2.200,3 novos casos por 100 mil habitantes.

Tabela 4 – Estatísticas descritivas da variável incidência por 100 mil pessoas

Mínimo	1º Quartil	Mediana	Média	3º Quartil	Máximo
0,009193	0,016147	0,022003	0,023963	0,027306	0,068889

A Figura 2 demonstra o histograma da variável incidência por 100 mil habitantes e a curva da distribuição beta. Pode-se ver que a distribuição beta apresentou um bom ajuste para a variável resposta.

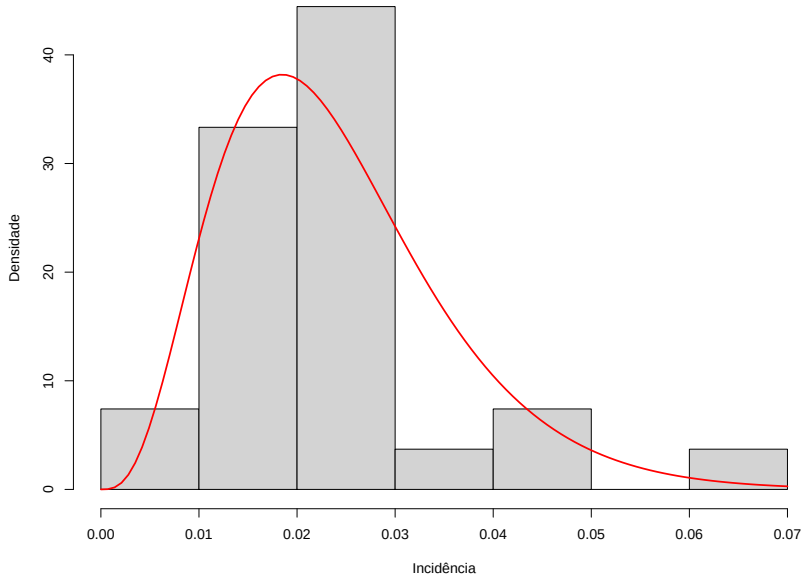

Figura 2 – Histograma da variável incidência por 100 mil pessoas e a curva da distribuição beta.

Além da análise descritiva da variável incidência, faz-se necessário verificar a relação entre a variável incidência e as demais covariáveis, como variáveis epidemiológicas e socioeconômicas. Nas Figuras 3 e 4 temos gráficos com informações da dispersão e correlação para os dados.

Ao analisar a Figura 3 é possível notar que a relação entre a variável incidência com o tempo da pandemia atinge uma correlação negativa fraca de -0,387, equanto que a variável incidência em relação a variável de pessoas com idade acima de 60 anos $(\%)$, apresenta uma correlação negativa forte de -0,755. Em relação aos gráficos de dispersão, nota-se que o gráfico da variável incidência em relação ao tempo da pandemia, apresenta uma dispersão linear negativa, indicando que quanto maior o tempo da pandemia, menor é a incidência.

Cabe destacar que a relação entre a variável incidência e a taxa de letalidade em $(\%)$ apresenta uma dispersão negativa, isso significa que quanto maior a taxa de letalidade, menor é a incidência, pois com a mortalidade alta terá menos pessoas para ser infectadas. Outra observação importante é a relação entre a incidência por 100 mil habitantes e leitos de UTI adulto por 10 mil habitantes, neste caso ocorre uma tendência linear negativa, indicando que a medida que a incidência aumenta os leitos de UTI diminuem.

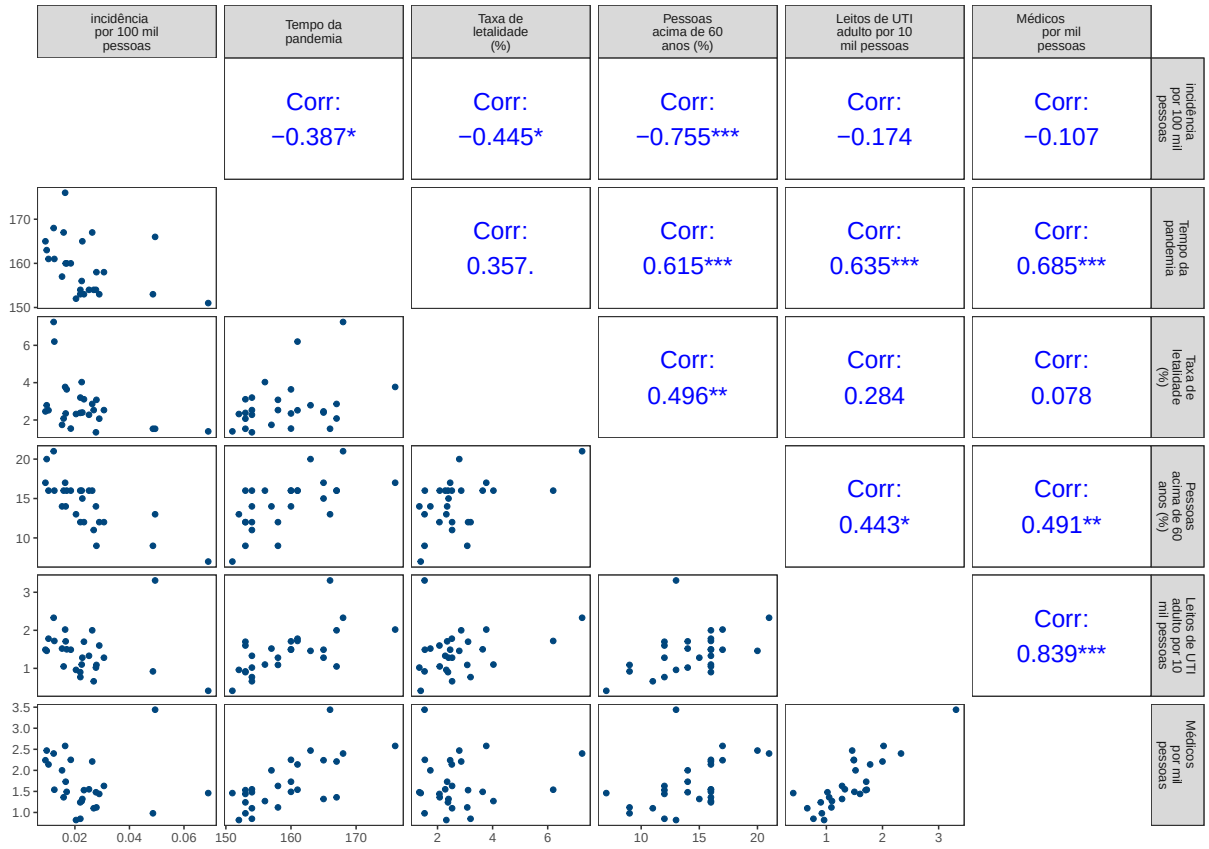

Figura 3 – Gráfico de dispersão e correlações da variável incidência sobre as demais variáveis

Em relação a Figura 4 é possível notar que a relação entre incidência e pessoas sem acesso a rede de esgoto é linear positiva, indicando que pessoas sem acesso a rede de esgoto são mais vulneráveis, embora a correlação seja fraca, o mesmo ocorre para a relação entre incidência e pessoas que não possuem acesso a rede geral de água, ambas as relações demonstram que as pessoas sem acesso a água e esgoto são mais vulneráveis. Outra observação importante a destacar é a relação entre a incidência e pessoas em domicílio com adensamento excessivo, neste caso existe uma correlação linear positiva moderada com tendência positiva.

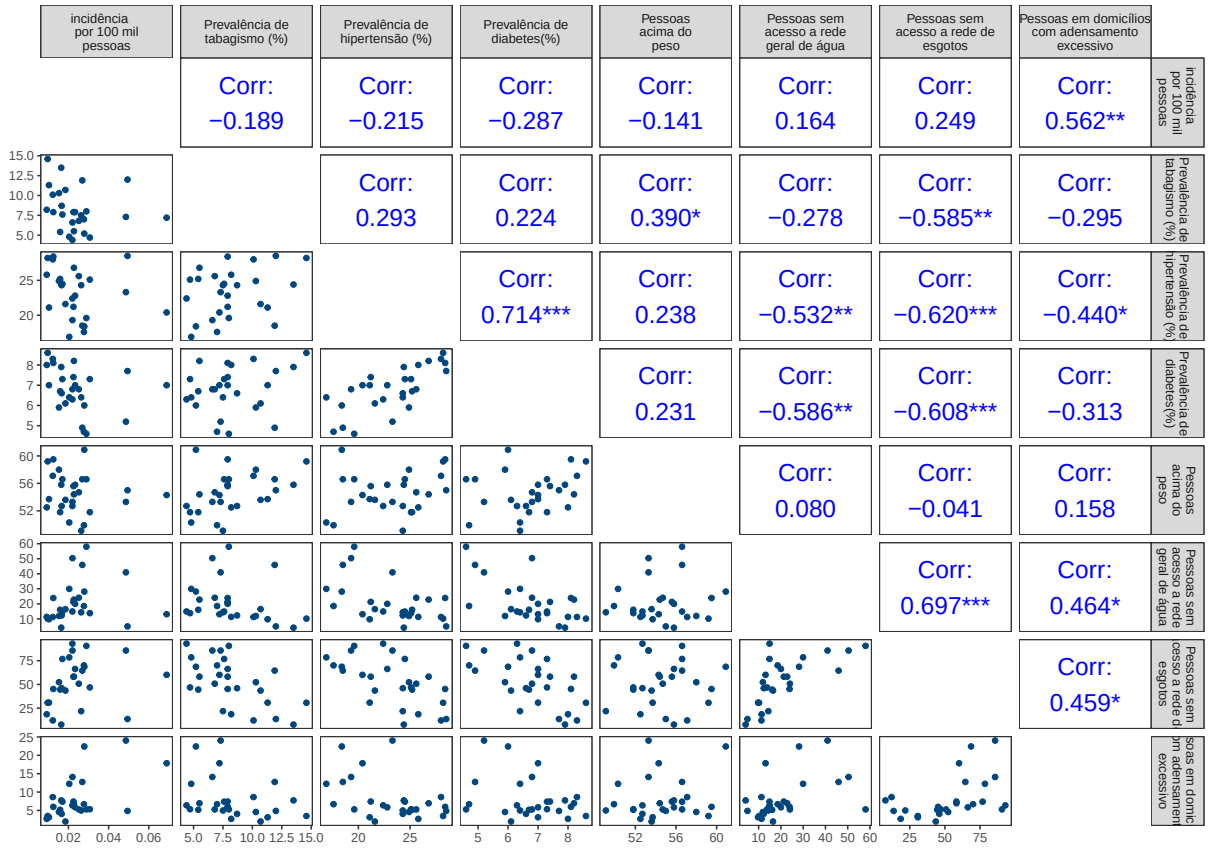

Figura 4 – Gráfico de dispersão e correlações da variável incidência sobre as variáveis com prevalência de condições crônicas, hábitos de vida e condições de moradia

O modelo de regressão beta será ajustado utilizando três tipos de funções de ligação: logito, probito e complemento log-log. Para a estimação dos parâmetros, foi utilizado o pacote `betareg` disponível no *software* `R` e com o intuito de encontrar um bom modelo, foi utilizando a função `StepBeta` do pacote `StepBeta` para selecionar modelos.

Com o intuito de encontrar um modelo que melhor se ajuste aos dados, foi util a Tabela 5 demonstra o melhor modelo selecionado pela função `StepBeta` utilizando a função de ligação logito. Considerando um nível de significância de 5%, pode-se ver que todos os valores-p são menores que 5%, ou seja, todos os parâmetros foram significativos.

As variáveis taxa de letalidade e tempo da pandemia apresentam uma estimativa negativa em relação a variável incidência, com isso duas variáveis diminuem a cada unidade no logito da média da incidência, enquanto que as variáveis médicos por mil habitantes, mortalidade por 100 mil habitantes e taxa de desocupação apresentam estimativas positivas, indicando que a cada aumento na unidade dessas variáveis ocorre um aumento no logito da média da incidência. Neste modelo, o pseudo R^2 foi igual a 95,13%, indicando a porcentagem que as variáveis explicativas do modelo refletem sobre a variável resposta incidência.

Tabela 5 – Ajuste do modelo de regressão beta com a função de ligação logito

	Estimativa	**Erro padrão**	**Valor Z**	**Valor-p**
(Intercepto)	-2,1555	0,8099	-2,6614	0,0078
Médicos por mil habitantes	0,1584	0,0506	3,1292	0,0018
Mortalidade por 100 mil habitantes	0,0182	0,0011	16,9545	<0,01
Taxa de letalidade	-0,3039	0,0211	-14,4060	<0,01
Taxa de desocupação	0,0235	0,0071	3,3034	0,0010
Temppo da pandemia	-0,0153	0,0058	-2,6416	0,0083
ϕ	5270	1435	-	-
Pseudo R^2		0,9513		
AIC		-245,4366		
BIC		-236,3657		

A Tabela 6 apresenta o melhor modelo selecionado pela função `StepBeta` do *software* `R`, utilizando como função de ligação a probito, nota-se que todas as variáveis apresentam um valor-p com nível de significância abaixo de 5%, indicando um melhor ajuste do modelo, em relação aos erros-padrão do modelo é possível verificar que o menor erro-padrão é o da variável PIB per capita, isso mostra que esta variável está mais próxima da reta de regressão ajustada, enquanto que neste modelo o maior erro-padrão encontra-se na variável índice de Gini, que mede o grau de desigualdade social, em relação as estimativas do modelo descrito na Tabela 6, nota-se que sete estimativas são negativas com exceção do intercepto, isso demonstra que essas variáveis diminuem a cada unidade no probito da média da variável incidência, um exemplo é a variável Leitos de UTI adulto por 10 mil habitantes. É importante destacar que neste modelo o Pseudo R^2 é igual a 98,33% de explicação das variáveis explicativas em relação a variável resposta incidência.

A Tabela 7 demonstra os dados ajustados do modelo de regressão beta tendo como função de ligação a complemento log-log, verifica-se que o valor-p de todas as variáveis estão abaixo ou igual a 5% indicando que as variáveis estão bem ajustadas a variável resposta incidência, em relação ao erro-padrão é possível verificar que os menores valores são das variáveis Mortalidade por 100 mil habitantes e Tempo da pandemia, essas variáveis demostram que estão mais próximas da reta de regressão ajustada.

Em relação as estimativas do modelo descrito na Tabela 7 é possível verificar que existem três variáveis com estimativas negativas, indicando que essas variáveis diminuem a cada unidade do complemento log-log da média da incidência, um exemplo desse comportamento é a variável Prevalência de diabetes. Outro fator a se notar é o Pseudo R^2, que verifica o quanto as covariáveis explicam a variável resposta incidência, neste caso o Pseudo R^2 atinge 95,98%.

Tabela 6 – Ajuste do modelo de regressão beta com a função de ligação probito

	Estimativa	**Erro padrão**	**Valor Z**	**Valor-p**
(Intercepto)	-2,2476	0,1043	-21,5437	<0,01
Pessoas acima de 60 anos	-0,0084	0,0035	-2,3682	0,0179
Mortalidade por 100 mil habitantes	0,0057	0,0005	10,8668	<0,01
Taxa de letalidade	-0,0874	0,0088	-9,9527	<0,01
Abaixo da linha da extrema pobreza	-0,0361	0,0052	-6,9804	<0,01
Rendimento médio mensal da população	0,0005	0,0001	7,6746	<0,01
Pessoas sem escolaridade	0,0194	0,0028	6,8027	<0,01
PIB per capita	<-0,01	<0,01	-4,7513	<0,01
Prevalência de tabagismo	-0,0225	0,0043	-5,1910	<0,01
Abaixo da linha da pobreza	0,0245	0,0035	6,9574	<0,01
Índice de Gini	-0,9959	0,2931	-3,3973	0,0007
Leitos de UTI adulto por 10 mil habitantes	-0,0641	0,0232	-2,7651	0,0057
Pessoas sem acesso a rede geral de água	0,0017	0,0005	3,2810	0,0010
ϕ	14373	3913	-	-
Pseudo R^2		0,9833		
AIC		-258,5373		
BIC		-240,3955		

Tabela 7 – Ajuste do modelo de regressão beta com a função de ligação complemento log-log

	Estimativas	**Erro padrão**	**Valor Z**	**Valor-p**
(Intercepto)	-2,2036	0,7481	-2,9457	0,0032
Médicos por mil habitantes	0,1818	0,0482	3,7715	0,0002
Mortalidade por 100 mil habitantes	0,0181	0,0010	18,1495	<0,01
Taxa de letalidade	-0,2860	0,0205	-13,9784	<0,01
Taxa de desocupação	0,0234	0,0065	3,6075	0,0003
Tempo da pandemia	-0,0140	0,0054	-2,5849	0,0097
Prevalência de diabetes	-0,0398	0,0201	-1,9771	0,0480
ϕ	6046	1646	-	-
Pseudo R^2		0,9598		
AIC		-247,1278		
BIC		-236,7611		

As Tabelas 5, 6 e 7 apresentam os três modelos ajustados com as três funções de ligação: *logit*, *probit* e *complementar log-log*, aparentemente o melhor modelo a se escolher seja o que está descrito na Tabela 5, embora apresente um pseudo R^2 menor que os demais, apresenta os menores erros-padrão de cada variável e os valores-p menor que 5%, além disso o *AIC* e *BIC* são menores em relação aos demais modelos.

A Tabela 8 apresenta as estimativas de MV e as estimativas de MV corrigidas do modelo utilizando a função de ligação logito, juntamente com as estimativas dos erros-padrão e as os valores-p do teste de Wald. Nota-se que as estimativas corrigidas para as variáveis regressoras não mudaram muito, porém a estimativa de ϕ corrigida apresentou uma diferença consubstancial. Além disso, todos os parâmetros relacionados às covariáveis foram significativos ao nível de significância de 5%.

Tabela 8 – Teste de Wald do modelo selecionado utilizando a função de ligação logito.

	$\hat{\theta}_{mv}$	$\hat{\theta}_{boot}$	Erro-padrão	Valor z	Valor-p
(Intercepto)	-2,1555	-2,1821	1,0723	-2,0349	0,0419
Médico por mil habitantes	0,1584	0,1564	0,0671	2,3325	0,0197
Mortalidade por 100 mil habitantes	0,0182	0,0182	0,0014	12,7974	<0,01
Taxa de letalidade	-0,3039	-0,3041	0,0279	-10,8891	<0,01
Taxa de desocupação	0,0235	0,0230	0,0094	2,4405	0,0147
Tempo da pandemia	-0,0153	-0,0151	0,0077	-1,9646	0,0495
ϕ	5270	2991,3806	815,0927	3,6700	0,0002

A Figura 5 demonstra o gráfico dos resíduos quantílicos. Nesse casso é possível verificar que os resíduos do modelo, utilizando a função de ligação logito, estão localizados dentro das bandas de confiança indicando que os resíduos não apresentam grandes afastamentos da normalidade.

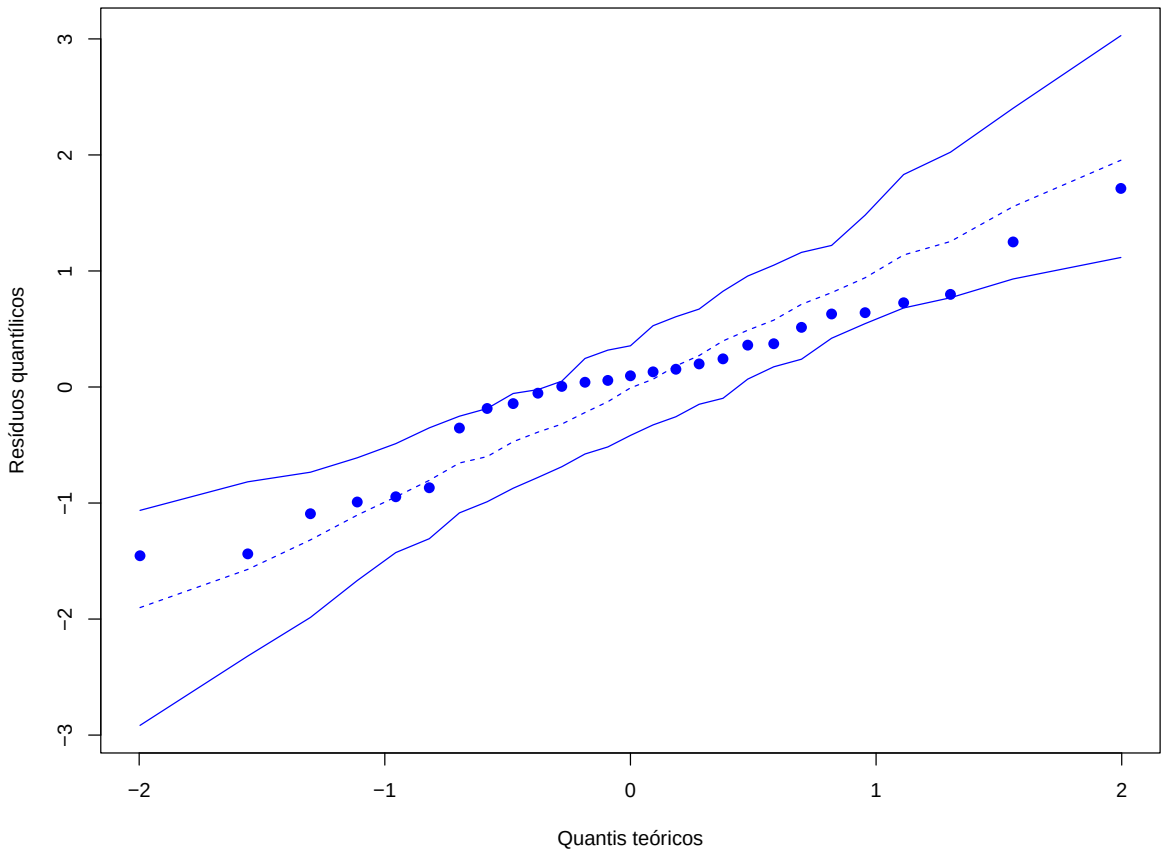

Figura 5 – Gráfico do envelope simulado dos quantis teóricos em relação resíduos.

A Figura 6 mostra os resíduos do modelo ajustado com a função de ligação logito em relação aos valores das observações. Nota-se que os resíduos possuem um comportamento aleatório, sem nenhum padrão evidente, dessa forma podemos afirmar que aparentemente os resíduos apresentam variância constante e não apresentam *outliers*.

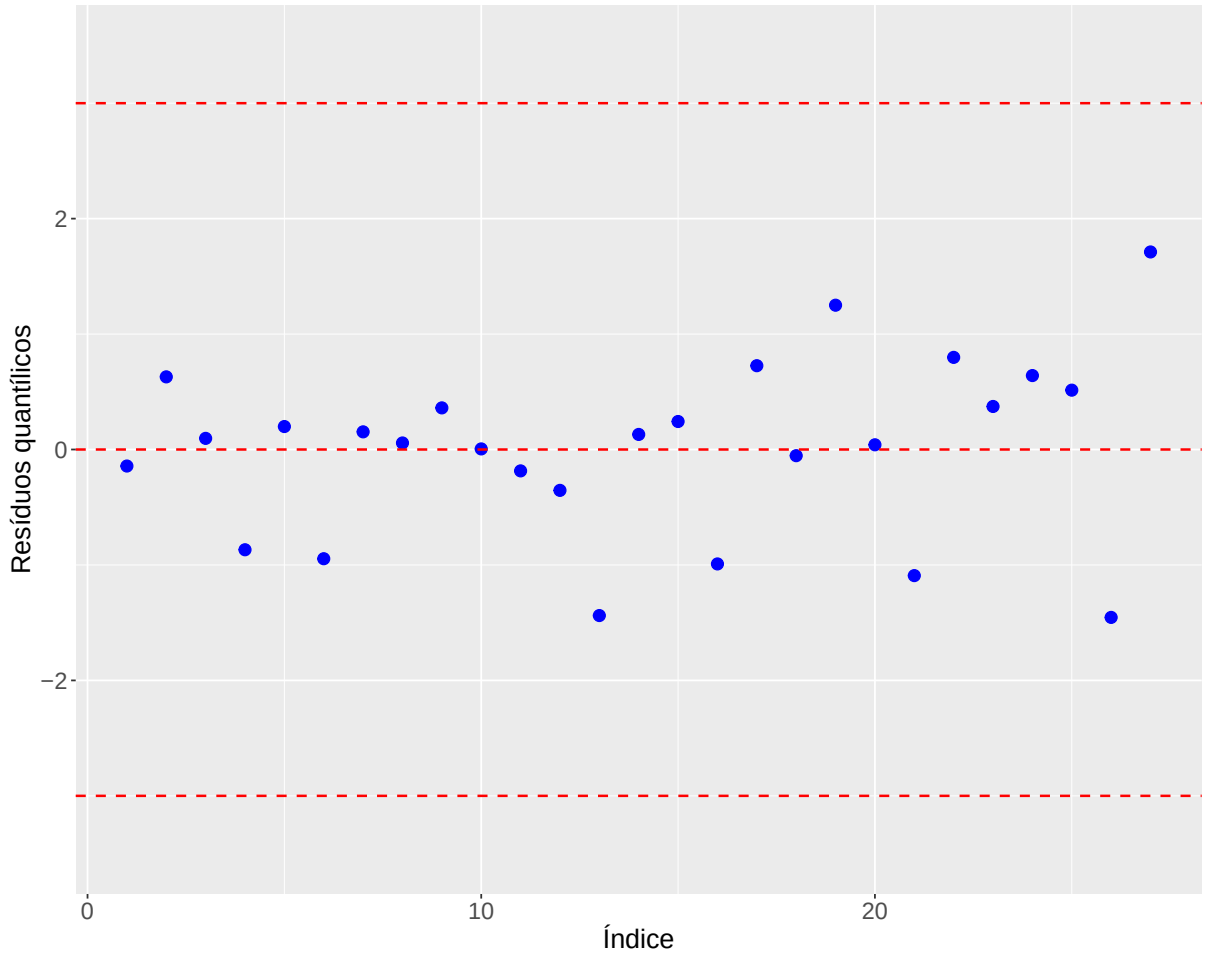

Figura 6 – Gráfico dos Resíduos versus Índices das observações do modelo ajustado.

Em relação a Figura 7 é possível verificar que o gráfico apresenta aleatoriedade dos resíduos do modelo em relação ao preditor linear. Neste caso é possível verificar que a função de ligação logito é adequada para o ajuste do modelo.

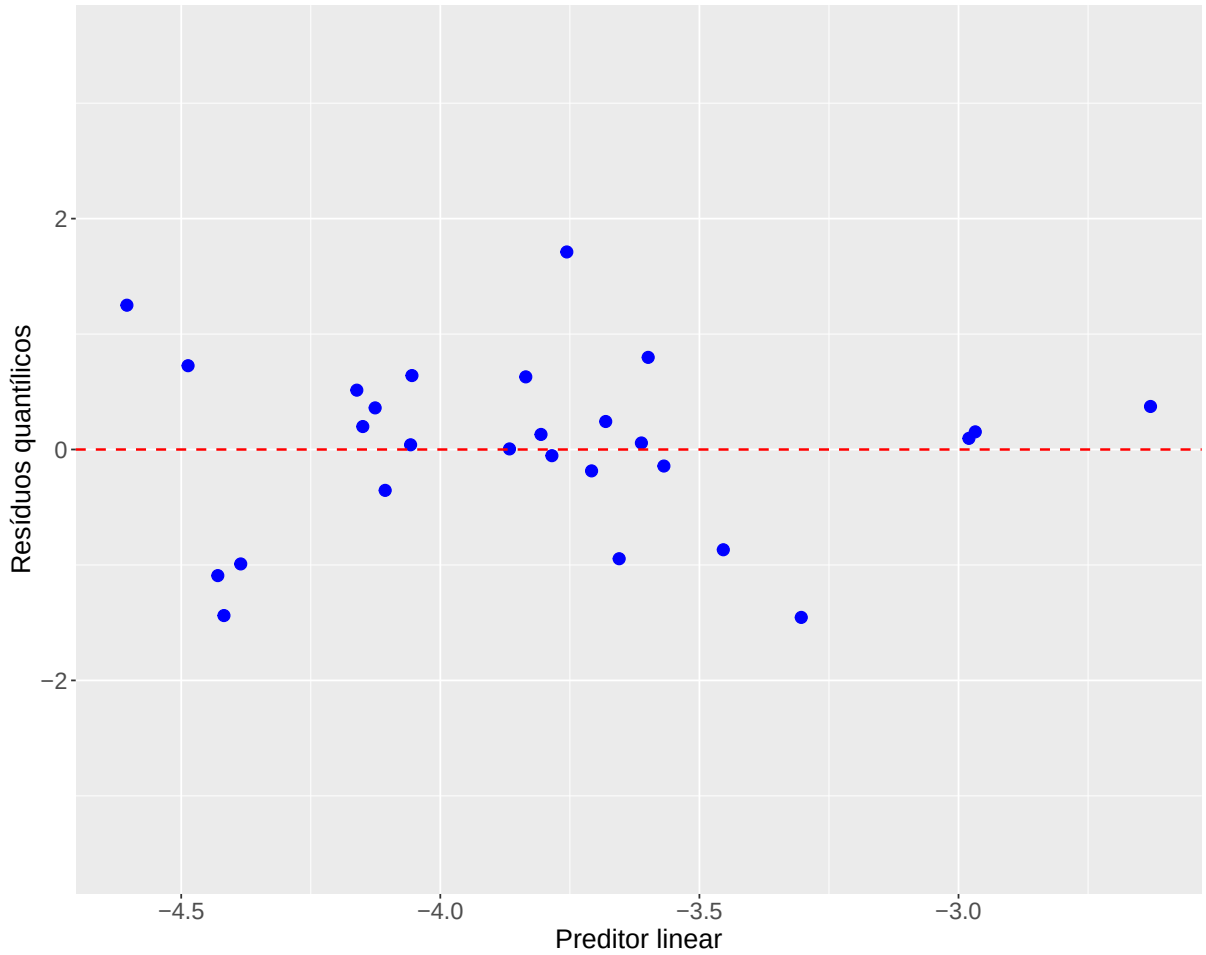

Figura 7 – Gráfico dos Resíduos versus Preditor do modelo ajustado.

Portanto, pode-se perceber que os resíduos do modelo não apresentaram problema, indicando que o modelo está bem ajustado aos dados.

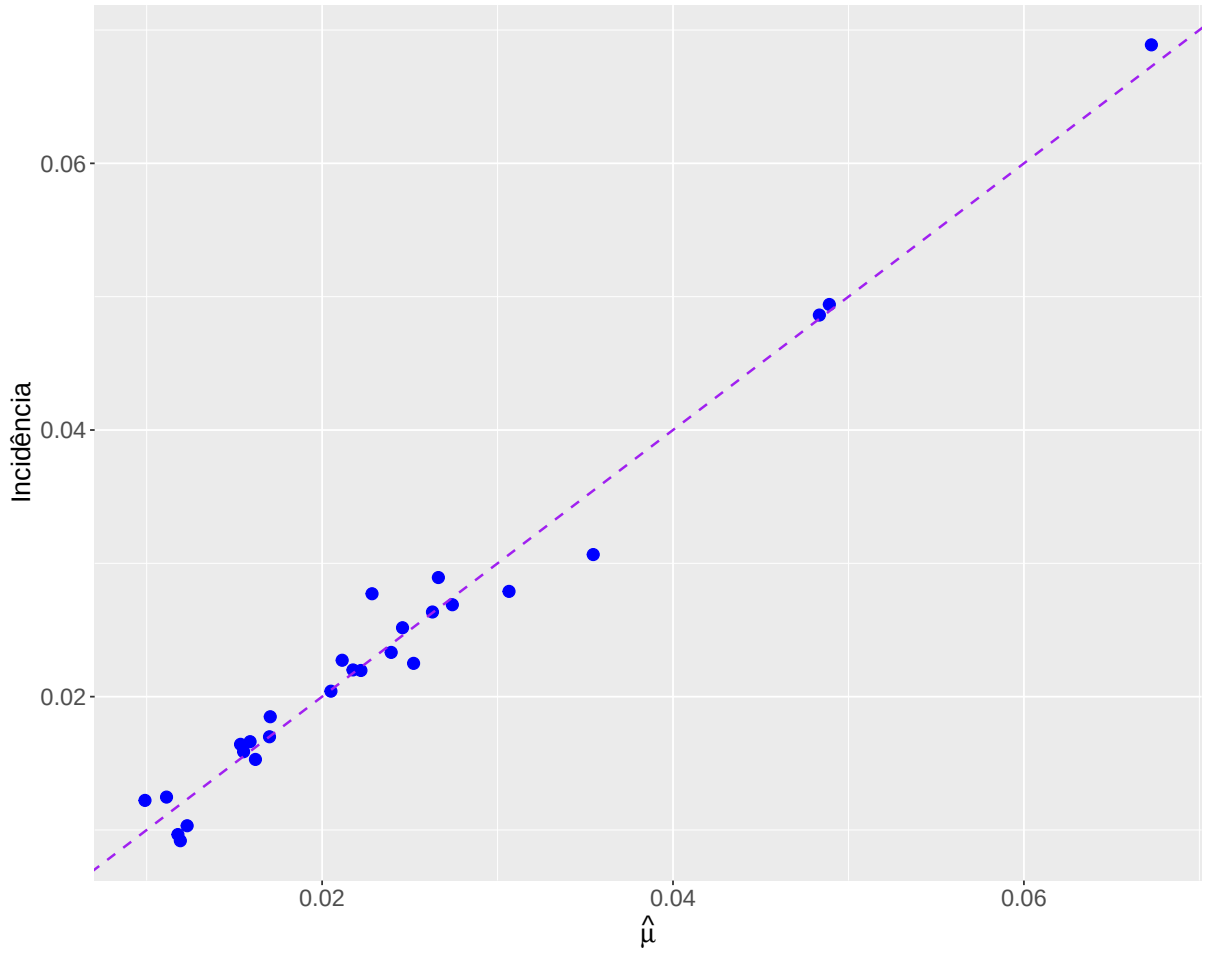

Figura 8 – Gráfico da Incidência versus $\hat{\mu}$ do modelo ajustado.

A Figura 8 demonstra a relação entre a incidência e o $\hat{\mu}$, neste caso é possível comprovar que $\hat{\mu}$ explica bem a variável resposta incidência, pois os dados estão próximos da reta $y = x$.

4 CONSIDERAÇÕES FINAIS

Neste trabalho foi abordado o modelo de regressão beta com dispersão variável, proposta por Ferrari e Cribari-Neto (2004), com aplicação em dados de Covid-19. Tendo como objetivo estudar os principais aspectos do modelo de regressão beta, além disso foi definido três objetivos específicos.

O primeiro objetivo específico foi apresentar os conceitos básicos da distribuição beta. Neste caso foi apresentado a distribuição beta e a distribuição beta reparametrizada, foi relatado que a distribuição beta reparametrizada tem como objetivo modelar a média da variável resposta. Além disso, fez-se necessário a definição do modelo de regressão beta reparametrizada em que foi estimado os parâmetros e a partir disso realizar a análise numérica para encontrar a média, desvio-padrão, curtose, coeficiente de assimetria e viés relativo. Dentro desses aspectos foi possível utilizar o critério de seleção *AIC* e *BIC*

O segundo objetivo específico deu-se na realização de um estudo de simulação para verificar o desempenho dos estimadores de MV para os parâmetros do modelo. Neste caso, a partir dos estimadores foi possível realizar a análise numérica.

O terceiro objetivo foi utilizar a correção do viés via *bootstrap* para os estimadores de MV, em que foram gerados 5 mil amostras de Monte Carlo e para cada amostra foi realizado a simulação de 1 mil amostras de *bootstrap*. Com isso, foi verificado que os estimadores de MV corrigidos, apresentam viéses menores em relação ao estimadores de MV.

O modelo de regressão beta com função de ligação logito foi o que a presentou o melhor ajuste, com base no AIC e BIC, em comparação com as funções de ligação probito e complemento log-log. Ao realizar o teste de Wald para o modelo escolhido, viu-se que todos os parâmetros foram significativos. Além disso, os resíduos do modelo não apresentaram *outliers*, afastamentos da normalidade e não constância da variância.

REFERÊNCIAS

AKAIKE, H. Information theory and an extension of the maximum likelihood principle. p. 267–281, 1973.

BAYER, F. M.; CRIBARI-NETO, F. Model selection criteria in beta regression with varying dispersion. **Communications in Statistics-Simulation and Computation**, Taylor & Francis, v. 46, n. 1, p. 729–746, 2017.

CASELLA, G.; BERGER, R. L. **Statistical inference**. [S.l.]: Cengage Learning, 2021.

DUNN, P. K.; SMYTH, G. K. Randomized quantile residuals. **Journal of Computational and graphical statistics**, Taylor & Francis, v. 5, n. 3, p. 236–244, 1996.

EFRON, B. Computers and the theory of statistics: thinking the unthinkable. **SIAM review**, SIAM, v. 21, n. 4, p. 460–480, 1979.

EFRON, B.; TIBSHIRANI, R. J. **An introduction to the bootstrap**. [S.l.]: CRC press, 1994.

FERRARI, S.; CRIBARI-NETO, F. Beta regression for modelling rates and proportions. **Journal of applied statistics**, Taylor & Francis, v. 31, n. 7, p. 799–815, 2004.

FIGUEIREDO, A. M. d.; FIGUEIREDO, D. C. M. M. d.; GOMES, L. B.; MASSUDA, A.; GIL-GARCÍA, E.; VIANNA, R. P. d. T.; DAPONTE, A. Determinantes sociais da saúde e infecção por covid-19 no brasil: uma análise da epidemia. **Revista Brasileira de Enfermagem**, SciELO Brasil, v. 73, 2020.

HAMMERSLEY, J. M.; HANDSCOMB, D. C. General principles of the monte carlo method. In: **Monte Carlo Methods**. [S.l.]: Springer, 1964. p. 50–75.

JÚNIOR, S. M.; VALADÃO, L. T.; VIEIRA, A. R. R.; MOURA, M. V. T. de. Análise de dados de vento para a região de botucatu-sp utilizando a distribuição beta. **Revista Brasileira de Agrometeorologia, Santa Maria**, v. 3, p. 129–132, 1995.

MORETTIN, P. A.; BUSSAB, W. O. **Estatística básica**. [S.l.]: Saraiva Educação SA, 2017.

OSPINA, P. L. E. **Regressão beta**. Tese (Doutorado) — Universidade de São Paulo, 2007.

OSPINA, R.; CRIBARI-NETO, F.; VASCONCELLOS, K. L. Improved point and interval estimation for a beta regression model. **Computational Statistics & Data Analysis**, Elsevier, v. 51, n. 2, p. 960–981, 2006.

PAWITAN, Y. **In all likelihood: statistical modelling and inference using likelihood**. [S.l.]: Oxford University Press, 2001.

SCHWARZ, G. Estimating the dimension of a model. **The annals of statistics**, JSTOR, p. 461–464, 1978.

SILVA, E. R. F. d. **Modelo de regressão beta modal**. 66f p. Dissertação (Mestrado) — Centro de Ciências Exatas e da Terra, Universidade Federal do Rio Grande do Norte, Natal, 2020.