UNIVERSIDADE ESTADUAL DA PARAÍBA
CAMPUS I
CENTRO DE CIÊNCIAS E TECNOLOGIA
DEPARTAMENTO DE ESTATÍSTICA
CURSO DE BACHARELADO EM ESTATÍSTICA

SUZIANE DE MELO ANDRADE

APLICAÇÃO DE MODELOS ADITIVOS GENERALIZADOS PARA LOCAÇÃO, ESCALA E FORMA NA ANÁLISE DA TAXA DE MORTALIDADE POR SUICÍDIO NO BRASIL

CAMPINA GRANDE - PB
2023

SUZIANE DE MELO ANDRADE

APLICAÇÃO DE MODELOS ADITIVOS GENERALIZADOS PARA LOCAÇÃO, ESCALA E FORMA NA ANÁLISE DA TAXA DE MORTALIDADE POR SUICÍDIO NO BRASIL

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.

Orientador: Profa. Dra. Ana Patrícia Bastos Peixoto
Coorientador: Profa. Dra. Wanessa Weridiana da Luz Freitas

CAMPINA GRANDE - PB
2023

A554a Andrade, Suziane de Melo.

Aplicação de modelos aditivos generalizados para locação, escala e forma na análise de taxa de mortalidade por suicídio no Brasil [manuscrito] / Suziane de Melo Andrade. - 2023.

42 p. : il. colorido.

Digitado.

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

"Orientação : Profa. Dra. Ana Patrícia Bastos Peixoto, Coordenação do Curso de Computação - CCT. "

"Coorientação: Profa. Dra. Wanessa Weridiana da Luz Freitas , UEPB - Universidade Estadual da Paraíba"

1. Inversa Gaussiana. 2. Taxa de suicídio. 3. Fatores socioeconômicos. I. Título

21. ed. CDD 519.5

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

Biblioteca Central BC/UEPB

SUZIANE DE MELO ANDRADE

APLICAÇÃO DE MODELOS ADITIVOS GENERALIZADOS PARA LOCAÇÃO, ESCALA E FORMA NA ANÁLISE DA TAXA DE MORTALIDADE POR SUICÍDIO NO BRASIL

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.

Trabalho aprovado em 30 de novembro de 2023 .

BANCA EXAMINADORA

Ana Patrícia Bastos Peixoto

Profa. Dra. Ana Patrícia Bastos Peixoto
Universidade Estadual da Paraíba (UEPB)

Elizabeth Mie Hashimoto

Profa. Dra. Elizabeth Mie Hashimoto
Universidade Tecnológica Federal do Paraná (UTFPR)

Fernanda Clotilde da Silva

Profa. Dra. Fernanda Clotilde da Silva
Universidade Estadual da Paraíba (UEPB)

Dedico este trabalho à minha amada mãe, a quem devo tudo. Seu amor incondicional, cuidado e incentivo são fundamentais na minha jornada. Ao meu pai (in memoriam), expresso minha profunda gratidão por todos os ensinamentos deixados; sua memória continuará viva em mim, inspirando-me a alcançar o melhor em cada passo.

AGRADECIMENTOS

A Deus, minha gratidão pela dádiva da vida, pelo alívio proporcionado em meio às minhas angústias, por iluminar meus passos e pelas forças que me dar para perseverar na jornada da vida.

À minha mãe e ao meu pai (*in memoriam*), as palavras não são suficientes para expressar o sentimento de gratidão e reconhecimento que nutro por vocês. À minha mãe, agradeço por todo amor, paciência e apoio ao longo de toda a trajetória acadêmica, tornando cada passo mais significativo. À memória eterna do meu pai, deixo registrado o apreço por tudo que representa em minha vida; seus ensinamentos foram cruciais para o meu desenvolvimento pessoal.

Aos meus irmãos, em especial a Verônica, Cristiane e Cristiano, por sempre estarem ao meu lado, incentivando o meu crescimento pessoal e profissional. Aos meus sobrinhos, por cada demonstração de carinho que compartilham comigo diariamente. São esses gestos afetuosos que iluminam meus dias e me ensinam constantemente sobre o verdadeiro significado do amor. À minha cunhada Izabela, pelas palavras de incentivo que constantemente me motivam a buscar o melhor em mim.

Aos meus amigos de sempre, Danielli, Gabrielli, Paloma, Dayana, Vanessa, Dhanny, Janaina, Rafaela, Wagner, Fernanda, Carol, Ricardo, Letícia, Renally, Rany e Rafael, agradeço por todo companheirismo, por estarem ao meu lado tornando os desafios mais leves, por cada momento proporcionando boas risadas. Sou grata pelo suporte e amor. Que possamos continuar construindo memórias e enfrentando juntos os altos e baixos que a vida nos reserva.

Às amizades que a UEPB me proporcionou, transformando o período da graduação em um ciclo marcante de afeto, trocas e aprendizados: Adelmo, Andrezza, Beatriz, Bruno, Biapino, Debora, Fagna, Felipe, Fernanda, Giullber, Gabriel, Isabele, Lucas, Zelma, Vicente, Lucas Oliveira. Aos que compartilharam comigo esses anos de jornada acadêmica, e também àqueles que cruzaram brevemente o meu caminho, agradeço por tornarem as noites universitárias mais leves e significativas.

Quero expressar minha gratidão à minha orientadora, Profa. Dra. Ana Patrícia Bastos Peixoto, pelo apoio, pela paciência diante das minhas dúvidas ao longo deste trabalho, e por todo incentivo, que foi fundamental. Agradeço também pela confiança depositada em mim.

Agradeço à minha coorientadora, Profa. Dra. Wanessa Weridiana da Luz Freitas, pelo apoio que me proporcionou ao longo deste percurso acadêmico. Sua disposição em auxiliar nas inúmeras vezes em que recorri a ela, sem hesitar em dedicar tempo e esforço para esclarecer minhas dúvidas. Sou imensamente grata por essa colaboração enriquecedora.

Agradeço a todos os professores, que fizeram parte da minha formação, por cada ensinamento e por terem dedicado tempo e esforço, contribuindo para o meu crescimento

tanto profissional quanto pessoal.

Agradeço, também, às professoras Profa. Dra. Elisabeth Mie Hashimoto e Profa. Dra. Fernanda Clotilde da Silva, por aceitarem o convite para integrar a banca examinadora deste trabalho. É muito importante contar com a colaboração de vocês.

Por fim, agradeço à Universidade Estadual da Paraíba e a toda a equipe de profissionais envolvidos. Durante minha trajetória nessa instituição, vivenciei momentos incríveis e desafiadores de muito crescimento e desenvolvimento.

RESUMO

De acordo com a Organização Mundial de Saúde estima-se que ocorram cerca de 800.000 mortes anuais por suicídio em todo o mundo. O Brasil é o oitavo país em número de suicídios, com média de 24 casos diários. Diante desse cenário, o suicídio é reconhecido como um sério problema de saúde pública. O presente estudo aborda a classe dos modelos aditivos generalizados para locação, escala e forma que possibilita o ajuste de uma ampla família de distribuições para a variável resposta. Além disso, essa metodologia oferece a capacidade de incorporar termos paramétricos e não paramétricos para as variáveis explicativas, proporcionando maior flexibilidade no ajuste do modelo. O objetivo do estudo é explorar a aplicação dos modelos aditivos generalizados para locação, escala e forma em um banco de dados referente às taxas de mortalidade por suicídio registrados nos Estados Brasileiros durante os anos de 2013 a 2017, afim de verificar a influência dos fatores socioeconômicos sobre essas taxas. Os dados utilizados foram obtidos a partir do Atlas de Desenvolvimento Humano no Brasil do IBGE. As análises foram conduzidas por meio do software R. Inicialmente, uma análise descritiva e a construção de um histograma revelaram uma assimetria positiva na variável resposta, com suporte no intervalo $(0, \infty)$. Diante desse perfil, tornou-se necessário o emprego de distribuições apropriadas para dados dessa natureza. A função fitDist(), disponível no pacote gamlss, foi empregada para ajustar diferentes distribuições, destacando-se a distribuição inversa Gaussiana como a mais adequada, considerando as características observadas na variável resposta. A seleção do modelo mais adequado foi realizada utilizando o critério de informação de Akaike generalizado. A adequabilidade do modelo foi avaliada por meio de análises gráficas dos resíduos, *worm plots* e o teste de Shapiro-Wilk. Os resultados obtidos indicaram que o modelo ajustado com a distribuição inversa Gaussiana foi o mais apropriado para modelar a taxa de mortalidade por suicídio. Este modelo final, possui dois parâmetros, μ e σ. As covariáveis, incluindo renda per capita, percentual de vulneráveis à pobreza, índice de Gini, índice de Theil e taxa de analfabetismo, demonstraram ser significativas ao nível de 5% de significância, evidenciando a influência desses fatores nas taxas de mortalidade por suicídio.

Palavras-chaves: GAMLSS; Inversa Gaussiana; Taxa de suicídio; Fatores socioeconômicos.

ABSTRACT

According to the World Health Organization, it is estimated that around 800,000 deaths by suicide occur every year around the world. Brazil is the eighth country in the number of suicides, with an average of 24 cases daily. Given this scenario, suicide is recognized as a serious public health problem. The present study addresses the class of generalized additive models for location, scale and shape that allow the fitting of a wide family of distributions for the response variable. Furthermore, this methodology offers the ability to incorporate parametric and non-parametric terms for explanatory variables, providing greater flexibility in model adjustment. The objective of the study is to explore the application of generalized additive models of location, scale and shape in a database referring to suicide mortality rates recorded in Brazilian states during the years 2013 to 2017 to verify the influence of socioeconomic factors on these rates. The data used were obtained from the IBGE Brazilian Atlas of Human Development. The analyses were carried out using the R software. Initially, a descriptive analysis and the construction of a histogram revealed a positive asymmetry in the response variable, supported by the interval $(0, \infty)$. Given this profile, it became necessary to use appropriate distributions for data of this nature. The fitDist() function, available in the gamlss package, was used to fit different distributions, highlighting the inverse Gaussian distribution as the most appropriate, considering the characteristics observed in the response variable. The selection of the most appropriate model was carried out using Akaike's generalized information criterion. The adequacy of the model was assessed using a graphical analysis of residuals, worm plots, and the Shapiro-Wilk test. The results obtained indicated that the model adjusted with the inverse Gaussian distribution was the most appropriate for modeling the suicide mortality rate. This final model has two parameters, μ and σ. The covariates, including per capita income, percentage of vulnerability to poverty, Gini index, Theil index, and illiteracy rate, were significant at the 5% significance level, demonstrating the influence of these factors on suicide mortality rates.

Keywords: GAMLSS; Inverse Gaussian; Suicide rate; Socioeconomic factors.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

O suicídio é considerado um grave problema de saúde pública mundial, que causa impactos na sociedade como um todo. De acordo com a Organização Mundial de Saúde (OMS), a cada ano, aproximadamente 800 mil vidas são perdidas devido ao suicídio em todo o mundo. No Brasil, a taxa de mortalidade por suicídio ultrapassou 6 casos por 100 mil habitantes em 2017 (Brasil, 2019). Em 10 anos, as taxas de mortalidade por suicídio aumentaram em 22%, passando de 4,9 mortes por 100 mil habitantes em 2008 para 6,02 em 2017 (Aguiar; Carvalho, 2019).

O ato do suicídio tem sido tradicionalmente associado a doenças mentais. No entanto, Durkheim (2000) apresenta uma perspectiva que vai além dessa visão. Ele argumenta que o suicídio não é meramente um ato individual determinado exclusivamente por fatores pessoais e relacionados apenas à psicologia. O suicídio tem causas objetivas que estão além do indivíduo, envolvendo a interação de fatores psicológicos, morais, sociais e culturais. O suicídio é considerado um reflexo do estado moral da sociedade.

Segundo Gonçalves, Gonçalves e Júnior (2011), fatores de natureza econômica desempenham um papel significativo no agravamento do bem-estar mental, aumentando assim a vulnerabilidade ao risco de suicídio. Além disso, Durkheim (2000) ressalta que o aumento da idade também exerce influência no cenário do suicídio. Pessoas idosas registram taxas mais elevadas de suicídio devido ao fato de que nesse estágio da vida os indivíduos enfrentam circunstâncias que tiram a sua vitalidade. Essas circunstâncias incluem o isolamento social, a experiência de perdas de entes queridos, o desemprego e desafios econômicos que geram aflições.

Os métodos estatísticos podem ser utilizados para uma melhor compreensão das taxas de suicídio. Ao longo das últimas décadas, houve um notável avanço no desenvolvimento de ferramentas para modelar diversos tipos de dados. Essa evolução tem sido impulsionada pela necessidade de lidar com situações complexas e muitas vezes mal compreendidas, onde modelos tradicionais podem não capturar adequadamente a variabilidade e as relações subjacentes nos dados.

Além disso, os modelos estatísticos são formulados com pressuposições rígidas sobre a distribuição dos dados e as relações entre as variáveis. Essas pressuposições, embora fossem convenientes para análises simples, muitas vezes não eram aplicáveis a situações do mundo real, levando a erros de especificação do modelo e a resultados imprecisos ou enviesados.

Os modelos normais lineares foram utilizados ao longo de muitos anos para descrever grande parte dos fenômenos aleatórios. Quando o fenômeno estudado apresentava uma variável resposta com problemas de linearidade, homocedasticidade da variância e normalidade, tentava-se alternativas de transformações para atender tais suposições (Paula, 2004).

Existem casos em que não é possível ajustar modelos lineares mesmo fazendo transformações na variável resposta, para alguns tipos de variáveis respostas a atribuição da distribuição normal não é aceita. Para atender essas limitações, foi proposto por Nelder e Wedderburn (1972) a classe dos modelos lineares gereralizados (MLG), ampliando a distribuição da variável resposta para a família exponencial de distribuições. Também foram propostos por Hastie e Tibshirani (1986, 1990) os modelos aditivos generalizados (MAG). Apesar da flexibilidade dos modelos normais lineares, o MLG e o MAG não possibilitam modelar uma variável em que sua distribuição não pertença a família exponencial.

Com o objetivo de suprir as limitações relacionadas aos modelos citados anteriormente, Rigby e Stasinopoulos (2005) desenvolveram a classe de modelos aditivos generalizados para locação, escala e forma (GAMLSS). Essa abordagem oferece uma solução flexível e abrangente para lidar com uma ampla variedade de distribuições da variável resposta, indo além das restrições impostas pela família exponencial de distribuições.

A principal inovação introduzida pelo GAMLSS foi a capacidade de modelar não apenas a média da distribuição da variável resposta, como é comum em muitos modelos estatísticos, mas também outros parâmetros importantes da distribuição. Isso permitiu uma maior adaptação dos modelos às características específicas dos dados observados.

Além disso, a estrutura do GAMLSS oferece a possibilidade de modelar a relação entre os parâmetros do modelo e as covariáveis de forma flexível. Isso significa que os efeitos das variáveis independentes podem ser incorporados de maneira paramétrica, não paramétrica, aditiva ou mesmo através de termos de efeitos aleatórios.

Nesse contexto, torna-se evidente que a classe GAMLSS representa uma ferramenta robusta para a modelagem estatística, proporcionando aos pesquisadores uma capacidade mais eficaz de lidar com a diversidade de distribuições e com situações em que os dados exibem comportamentos assimétricos, demandando a utilização de modelos mais complexos (Leite, 2019).

Com base nas informações apresentadas, este estudo tem como objetivo principal a utilização do GAMLSS. Através dessa abordagem, busca-se não apenas evidenciar a sua notável flexibilidade na modelagem de conjuntos de dados, mas também realizar uma análise dos fatores socioeconômicos que exercem influência sobre as taxas de mortalidade por suicídio nas unidades federativas do Brasil.

2 REVISÃO DA LITERATURA

Para uma melhor compreensão do assunto abordado, nesse Capítulo será introduzido conceitos relacionados a aplicação do GAMLSS. Além disso, serão abordadas informações referentes aos métodos de estimação, inferência, critério de seleção do modelo e análise de resíduos em GAMLSS.

2.1 Modelos aditivos generalizados para locação, escala e forma

Os modelos aditivos generalizados para locação, escala e forma (GAMLSS), são uma classe de modelos de regressão univariada desenvolvida por Rigby e Stasinopoulos (2005). Consiste em um método de modelagem que admite o ajuste de um modelo sem a exigência de que a distribuição de uma variável resposta pertença apenas a família exponencial, podendo pertencer então a uma família de distribuições gerais. No GAMLSS é permitido que não só a média, mas todos os parâmetros da distribuição da variável resposta Y sejam modelados em função das variáveis explicativas. Além disso, podem ser incluídos ao preditor funções paramétricas e não-paramétricas de suavização, termos aditivos ou de efeitos aleatórios.

No GAMLSS são utilizados modelos aditivos para a modelagem de p parâmetros $\boldsymbol{\theta}^T = (\theta_1, \theta_2, ..., \theta_p)$, de uma função densidade de probabilidade $f(y_i|\theta^i)$. Considera-se que y_i com $i = 1, 2, ..., n$, são condicionais independentes a $\boldsymbol{\theta}^i$, com função densidade de probabilidade $f(y_i|\theta^i)$, onde $\boldsymbol{\theta}^{iT} = (\theta_{i1}, \theta_{i2}, ..., \theta_{ip})$ é um vetor de p parâmetros relacionados a variáveis explanatórias e efeitos aleatórios. Ressalta-se que quando os valores assumidos pelas covariáveis são estocásticos ou as observações y_i dependem de seus valores passados então $f(y_i|\theta^i)$ é interpretada como sendo condicional a esses valores.

Seja $\boldsymbol{y}^T = (y_1, y_2, ..., y_n)$ o vetor das observações da variável resposta, considerando também para $k = 1, 2, ..., p$, uma função de ligação monótona $g_k(.)$ com o k-ésimo parâmetro θ_k sendo relacionado com as variáveis explicativas e efeitos aleatórios por meio de um modelo aditivo dado por

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = \boldsymbol{X}_k\boldsymbol{\beta}_k + \sum_{j=1}^{J_k} \boldsymbol{Z}_{jk}\gamma_{jk}, \tag{2.1}$$

em que $\boldsymbol{\theta}_k$ e $\boldsymbol{\eta}_k$ são vetores de tamanho $(n \times 1)$, por exemplo, $\boldsymbol{\theta}_k^T = (\theta_{1k}, \theta_{2k}, ..., \theta_{nk})$, $\boldsymbol{\beta}_k^T = (\beta_{1k}, \beta_{2k}, ..., \beta_{J_k'k})$ é um vetor de parâmetro de tamanho $J_k^{'}$, $\boldsymbol{X}_k$ e $\boldsymbol{Z}_{jk}$ são matrizes de planejamento (covariáveis) conhecidas e de ordens $(n \times J_k^{'})$ e $(n \times q_{jk})$, respectivamente. Sendo $\boldsymbol{\gamma}_{jk}$ definida como uma variável aleatória com dimensão q_{jk}-dimensional. O modelo (2.1) é denominado de GAMLSS (Rigby; Stasinopoulos, 2005).

Os vetores $\boldsymbol{\gamma}_{jk}$, para $J = 1, 2, ..., J_k$, podem ser combinados em um único vetor $\boldsymbol{\gamma}_k$ e em uma única matriz de projeto $\boldsymbol{Z}_k$, No entanto, a formulação proposta em (2.1) acaba sendo mais apropriada por permitir que combinações de diferentes tipos de termos

aditivos e/ou efeitos aleatórios sejam facilmente incorporados ao modelo e por facilitar o uso dos algoritmos de retroajuste, conhecido como *backfitting* (Rigby; Stasinopoulos, 2005). Quando $j_k = 0$, para $k = 1, 2, ..., p$, não existe termos aditivos associados aos parâmetros da distribuição, sendo assim, o modelo (2.1) se reduz a um modelo linear completamente paramétrico dado por

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = \boldsymbol{X}_k\boldsymbol{\beta}_k, \tag{2.2}$$

em que $\boldsymbol{\theta}_k$ e $\boldsymbol{\eta}_k$ são vetores de tamanho $(n \times 1)$, $\boldsymbol{\beta}_k^T = (\beta_{1k}, \beta_{2k}, ..., \beta_{J_k'k})$ é um vetor de parâmetro de tamanho J_k', e $\boldsymbol{X}_k$ é uma matriz de planejamento (covariáveis) conhecida e de ordem $(n \times J_k')$.

Se $\boldsymbol{Z}_{jk} = \boldsymbol{I}_n$, em que $\boldsymbol{I}_n$ é uma matriz identidade de ordem $n \times n$, e $\gamma jk = \boldsymbol{h}_{jk} = h_{jk}(xjk)$ para todas as combinações de j e k no modelo (2.1), tem-se

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = \boldsymbol{X}_k\boldsymbol{\beta}_k + \sum_{j=1}^{J_k} \boldsymbol{h}_{jk}(\mathrm{x}_{jk}), \tag{2.3}$$

onde x_{jk}, para $j = 1, ..., J_k$ e $k = 1, ..., p$, são vetores de tamanho n. A função h_{jk} é uma função desconhecida da variável explicativa x_{jk} e $\boldsymbol{h}_{jk} = h_{jk}(\mathrm{x}_{jk})$, é um vetor que avalia h_{jk} em x_{jk}. Sendo assim, assume-se que vetores x_{jk} são conhecidos e o modelo apresentado na Equação (2.3) é denominado de GAMLSS aditivo semi-paramétrico linear (Rigby; Stasinopoulos, 2005).

A Equação (2.3) pode ser ampliada para que sejam incluídos termos não-lineares na modelagem dos k parâmetros da distribuição, do seguinte modo

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = h_k(\boldsymbol{X}_k, \boldsymbol{\beta}_k) + \sum_{j=1}^{J_k} h_{jk}(\mathrm{x}_{jk}), \tag{2.4}$$

no qual h_k para $k = 1, 2, ..., p$ são funções não-lineares e $\boldsymbol{X}_k$ é uma matriz de covariáveis conhecida de ordem $n \times J_k''$. O modelo (2.4) é denominado de GAMLSS aditivo semi-paramétrico não-linear. No caso de $J_k = 0$, o modelo (2.4) reduz-se a um modelo GAMLSS paramétrico não linear, dado por

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = h_k(\boldsymbol{X}_k, \boldsymbol{\beta}_k). \tag{2.5}$$

Quando $h_k(\boldsymbol{X}_k, \boldsymbol{\beta}_k) = \boldsymbol{X}_k^T, \boldsymbol{\beta}_k$, para $i = 1, 2, ..., n$ e $k = 1, 2, ..., p$, a Equação (2.5) se reduz ao modelo paramétrico linear (2.2). Note que alguns dos termos de $h_k(\boldsymbol{X}_k, \boldsymbol{\beta}_k)$ podem ser lineares, resultando então em um GAMLSS com a combinação de termos paramétricos lineares e não lineares (Rigby; Stasinopoulos, 2006).

É comum encontrar trabalhos que atribuem quatro parâmetros ($p = 4$), que são normalmente caracterizados por locação (μ), escala (σ), assimetria (ν) e curtose (τ). Vale ressaltar que a do modelo GAMLSS, tem a opção de fazer modelagem com mais de quatro parâmetros. Na literatura, os parâmetros populacionais $\mu = \theta_1$ e $\sigma = \theta_2$ no modelo (2.1),

são definidos por parâmetros de locação e escala, e os demais parâmetros $\nu = \theta_3$ e $\tau = \theta_4$ são definidos como parâmetros de forma. Dessa maneira, define-se:

$$\begin{aligned} g_1(\boldsymbol{\mu}) &= \boldsymbol{\eta}_1 = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \textstyle\sum_{j=1}^{J_1} \boldsymbol{Z}_{j1}\boldsymbol{\gamma}_{j1}, \\ g_2(\boldsymbol{\sigma}) &= \boldsymbol{\eta}_2 = \boldsymbol{X}_2\boldsymbol{\beta}_2 + \textstyle\sum_{j=1}^{J_2} \boldsymbol{Z}_{j2}\boldsymbol{\gamma}_{j2}, \\ g_3(\boldsymbol{\nu}) &= \boldsymbol{\eta}_3 = \boldsymbol{X}_3\boldsymbol{\beta}_3 + \textstyle\sum_{j=1}^{J_3} \boldsymbol{Z}_{j3}\boldsymbol{\gamma}_{j3}, \\ g_4(\boldsymbol{\tau}) &= \boldsymbol{\eta}_4 = \boldsymbol{X}_4\boldsymbol{\beta}_4 + \textstyle\sum_{j=1}^{J_4} \boldsymbol{Z}_{j4}\boldsymbol{\gamma}_{j4}. \end{aligned} \tag{2.6}$$

2.1.1 Estimação

Para o ajuste de componentes aditivos dentro da estrutura GAMLSS, existe dois fatores importantes, sendo eles o algoritmo *backfitting* (Hastie; Tibshirani, 1986) e o fato de que as penalizações quadráticas na função de verossimilhança resultam do pressuposto de que os efeitos aleatórios no preditor linear seguem distribuição normal. A partir disto, o processo de estimação do modelo utiliza matrizes de encolhimento (suavização) relacionado ao algoritmo de *backfitting*.

Considerando que no modelo (2.1) os efeitos aleatórios $\boldsymbol{\gamma}_{jk}$ são independentes entre si e seguem distribuição normal com $\boldsymbol{\gamma}_{jk} \sim N_{qjk}(\boldsymbol{0}, \boldsymbol{G}_{jk}^{-1})$, no qual $\boldsymbol{G}_{jk}^{-1}$ é a inversa (generalizada) com ordem $q_{jk} \times q_{jk}$ da matriz simétrica $\boldsymbol{G}_{jk} = \boldsymbol{G}_{jk}(\boldsymbol{\lambda}_{jk})$. Esta matriz pode depender de um vetor de hiperparâmetros $\boldsymbol{\lambda}_{jk}$ e, se $\boldsymbol{G}_{jk}$ for singular, considera-se para $\boldsymbol{\gamma}_{jk}$ uma função de densidade imprópria proporcional à $\exp\left(-\frac{1}{2}\boldsymbol{\gamma}_{jk}^T\boldsymbol{G}_{jk}\boldsymbol{\gamma}_{jk}\right)$. Para simplificar a notação $\boldsymbol{G}_{jk}(\boldsymbol{\lambda}_{jk})$ é referido apenas como $\boldsymbol{G}_{jk}$, embora ainda continue existindo a dependência de $\boldsymbol{G}_{jk}$ em hiperparâmetros $\boldsymbol{\lambda}_{jk}$.

O pressuposto de independência entre os diferentes vetores de efeitos aleatórios $\boldsymbol{\gamma}_{jk}$ é importante dentro do contexto da estrutura GAMLSS. No caso de dois ou mais vetores de efeitos aleatórios não serem independentes para um determinado k, eles podem então ser combinados em um único vetor de efeitos aleatórios. Da mesma forma, as matrizes de planejamentos $\boldsymbol{Z}_{jk}$ podem ser transformadas em uma única matriz, satisfazendo a condição de independência.

Em um estudo conduzido por Rigby e Stasinopoulos (2005), foram utilizados argumentos Bayesianos empíricos que demonstraram a equivalência entre o método de estimação máxima a posteriori (MAP) para os vetores de parâmetro $\boldsymbol{\beta}_k$ e os termos de efeitos aleatórios $\boldsymbol{\gamma}_{jk}$ (para valores fixos de suavização ou hiperparâmetros $\boldsymbol{\lambda}_{jk}$), onde $j = 1, 2, ..., J_k$ e $k = 1, 2, ..., p$, com a estimação por máxima verossimilhança penalizada.

Sendo assim, para os valores fixos de $\boldsymbol{\lambda}_{jk}$, os $\boldsymbol{\beta}_k$'s e os $\boldsymbol{\gamma}_{jk}$'s são estimados na estrutura GAMLSS, através da maximização da função de verossimilhança penalizada ℓ_p, dada por

$$\ell_p = \ell - \frac{1}{2}\sum_{k=1}^{p}\sum_{j=1}^{J_k} \boldsymbol{\gamma}_{jk}^T\boldsymbol{G}_{jk}\boldsymbol{\gamma}_{jk}, \tag{2.7}$$

no qual $\ell = \sum_{i=1}^{n} \log\{f(y_i \mid \boldsymbol{\theta}^i)\}$ é a função de log-verossimilhança dos dados fornecidos a $\boldsymbol{\theta}_i$ para $i = 1, 2, ..., n$. Isto é equivalente a maximizar a verossimilhança estendida ou hierárquica defina por

$$\ell_h = \ell_p + \frac{1}{2}\sum_{k=1}^{p}\sum_{j=1}^{J_k}\{\log \mid \boldsymbol{G}_{jk} \mid -q_{jk}\log(2\pi)\}.$$

Para mais detalhes ver Lee e Nelder (1996) e Pawitan (2001).

Rigby e Stasinopoulos (2005) demonstram que a maximização de (2.7) aplicada aos resíduos parciais $\boldsymbol{\varepsilon}_{jk}$ (ver Rigby e Stasinopoulos (2005)), para atualizar a estimativa do preditor aditivo $\boldsymbol{Z}_{jk}\lambda_{jk}$, resulta na matriz de encolhimento (suavização) $\boldsymbol{S}_{jk}$, dada por

$$\boldsymbol{S}_{jk} = \boldsymbol{Z}_{jk}(\boldsymbol{Z}_{jk}^T\boldsymbol{W}_{kk}\boldsymbol{Z}_{jk} + \boldsymbol{G}_{jk})^{-1}\boldsymbol{Z}_{jk}^T\boldsymbol{W}_{kk}, \tag{2.8}$$

para $j = 1, 2, ..., J_k$ e $k = 1, 2, ..., p$, na qual $\boldsymbol{W}_{kk}$ é uma matriz diagonal de pesos iterativos.

Diferentes formas de $\boldsymbol{Z}_{jk}$ e $\boldsymbol{G}_{jk}$ correspondem a diferentes tipos de termos aditivos no preditor linear $\boldsymbol{\eta}_k$ para $k = 1, 2, ..., p$. $\boldsymbol{G}_{jk}$ é normalmente uma matriz de baixa ordem, para os termos de efeitos aleatórios, considerando que para um termo de suavização *spline cúbico*, tem-se $\boldsymbol{\gamma}_{jk} = \boldsymbol{h}_{jk}$, $\boldsymbol{Z}_{jk} = \boldsymbol{I}_n$ e $\boldsymbol{G}_{jk} = \lambda_{jk}\boldsymbol{K}_{jk}$ onde $\boldsymbol{K}_{jk}$ é uma matriz estruturada. Qualquer um dos casos permitem a atualização de $\boldsymbol{Z}_{jk}\boldsymbol{\gamma}_{jk}$.

2.1.2 Preditor linear

No GAMLSS (2.1), existem componentes paramétricos representados como $\boldsymbol{X}_k\boldsymbol{\beta}_k$, para $k = 1, 2, ..., p$, e componentes aditivos denotados por $\boldsymbol{Z}_{jk}\boldsymbol{\gamma}_{jk}$, onde $j = 1, 2, ..., J_k$, ambos integrados no preditor linear. A seguir, serão destacadas algumas observações pertinentes relacionadas a cada um desses elementos.

Termos paramétricos: o componente pode permitir a inclusão de diversos termos, tais como termos lineares, termos de interação, fatores, polinômios, polinômios fracionários (Royston; Altman, 1994) e até mesmo polinômios segmentados com nós fixos, todos adaptáveis às variáveis exploratórias em questão.

Termos aditivos: os termos aditivos representados por $\boldsymbol{Z}_{jk}\boldsymbol{\gamma}_{jk}$ no modelo (2.1), têm a capacidade de modelar uma série de termos, incluindo suavização, efeitos aleatórios e componentes relevantes para a análise de séries temporais, como, por exemplo, passeios aleatórios. Entre os termos aditivos mais comumente utilizados destacam-se os *spline* cúbicos, *spline* de penalização, polinômios fracionários e *loess*. Para uma compreensão mais aprofundada desses termos aditivos, recomenda-se a leitura dos trabalhos de Cleveland, Grosse e Shyu (1992) e Hastie e Tibshirani (1990). Na Tabela 1 estão descritos alguns dos termos aditivos disponíveis no GAMLSS.

Tabela 1 – Termos aditivos implementados no pacote `gamlss`.

Termo aditivo	**Nomenclatura**
Cubic splines	cs(), scs()
Decision tress	tr()
Fractional and power polynomials	fp(), pp()
Free knots (break points)	fk()
Loess	lo()
Neural networks	nn()
Nonlinear flt	nl()
P-splines	pb(), pb0(), ps()
P-splines cyclic	pbc(), cy()
P-splines monotonic	pbm()
P-splines shrinking to zero	pbz()
P-splines varying coefflcient	pvc()
Penalized categorical	pcat()
Random effects	random(), re()
Ridge regression	ri()
Simon Wood's gam	ga()
Stephen Milborrow's earth	ma()

Fonte: Adaptado de Rigby et al. (2019).

2.1.3 Inferência

Na modelagem GAMLSS é preciso utilizar ferramentas inferenciais que tenham a capacidade para atender a variedade de distribuições e termos no modelo, funções de ligação que podem ser assumidas por cada parâmetro da distribuição e ainda diferentes conjuntos de variáveis explicativas. Stasinopoulos et al. (2017) atribuem métodos inferenciais que são utilizados para resolver questões relacionadas a inferência, sendo elas a inferência baseada na função de verossimilhança e o *bootstrapping*. Para o caso de um GAMLSS paramétrico, seja $\boldsymbol{\theta}$ o vetor de parâmetros do modelo, com os coeficientes lineares para μ, σ, ν ou τ, isto é, $(\beta_1, \beta_2, \beta_3, \beta_4)$. De acordo com a teoria, assintoticamente, tem-se

$$\hat{\boldsymbol{\theta}} \sim N\left(\boldsymbol{\theta}_T, \boldsymbol{i}(\boldsymbol{\theta}_T)^{-1}\right),$$

onde $\hat{\boldsymbol{\theta}}$ é o estimador de máxima verossimilhança e

$$\boldsymbol{i}(\boldsymbol{\theta}_T) = -E\left[\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\right]_{\boldsymbol{\theta}_T},$$

é a matriz de informação esperada de Fisher. Nos casos em que a matriz de informação esperada não pode ser obtida, utiliza-se a matriz de informação observada de Fisher definida por

$$\boldsymbol{I}(\boldsymbol{\theta}_T) = -\left[\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\right]_{\boldsymbol{\theta}_T}.$$

A estimativa da matriz de variâncias e covariâncias da distribuição assintótica de $\hat{\boldsymbol{\theta}}$ é aproximada por $\boldsymbol{I}(\boldsymbol{\theta}_T)^{-1}$ ao invés de $\boldsymbol{i}(\boldsymbol{\theta}_T)^{-1}$. Quando o $\boldsymbol{\theta}_T$ for desconhecido, será então substituído por $\hat{\boldsymbol{\theta}}$ nas matrizes de informação esperada e observadas, dadas por $\boldsymbol{i}(\hat{\boldsymbol{\theta}})$ e $\boldsymbol{I}(\hat{\boldsymbol{\theta}})$, respectivamente (Stasinopoulos et al., 2017).

Para os modelos GAMLSS paramétricos, é usada a seguinte distribuição assintótica para $\hat{\boldsymbol{\theta}}$

$$\hat{\boldsymbol{\theta}} \sim N\left(\boldsymbol{\theta}_T, \mathbf{I}(\hat{\theta})^{-1}\right).$$

Os erros padrões dos parâmetros estimados são obtidos através da raiz quadrada dos elementos da diagonal da matriz de variâncias e covariâncias relacionada a $\boldsymbol{\theta}$. Nos casos em que a obtenção da matriz de informação é complicada, uma alternativa sugerida por Aitkin et al. (2009) para obter o erro padrão de um particular $\hat{\boldsymbol{\beta}}$ é dada por

$$EP(\hat{\boldsymbol{\beta}}) \approx \frac{\mid \hat{\boldsymbol{\beta}} \mid}{\sqrt{\Delta GDEV}},$$

em que $\Delta GDEV$ é a diferença do desvio global obtida pela omissão das variáveis explicativas associadas ao coeficiente do modelo. Esse procedimento se justifica aproximando a estatística de teste de razão de verossimilhanças à estatística de Wald, para que seja testado a hipótese nula $\beta = 0$, em que, (Stasinopoulos et al., 2017)

$$\left[\frac{\hat{\boldsymbol{\beta}}}{EP(\hat{\boldsymbol{\beta}})}\right]^2 \approx \Delta GDEV.$$

2.1.4 Algoritmo

Para ajustar um GAMLSS, é recomendado o uso de dois algoritmos, quando trata-se de valores fixos de hiperparâmetros, com o objetivo de maximizar a função de verossimilhança penalizada (Rigby; Stasinopoulos, 2005). Esses algoritmos são implementados em liguagem R. O software R foi desenvolvido por Ross Ihaka e Robert Gentleman. É um ambiente integrado, destacando-se pela sua distribuição livre e pelo código fonte aberto, oferecendo amplas facilidades para a manipulação de dados, realização de cálculos, geração de gráficos e a modelagem estatística em geral (Ihaka; Gentleman, 1996).

O algoritmo CG (Cole e Green) é uma generalização do algoritmo de Cole e Green (1992) e faz o uso da primeira derivada e o valor esperado ou aproximado das derivadas de segunda ordem e das derivadas cruzadas da função log-verossimilhança em relação aos parâmetros da distribuição $\boldsymbol{\theta} = (\mu, \sigma, \nu, \tau)$ com quatro parâmetros.

Para muitas funções de densidade de probabilidade $f(y \mid \boldsymbol{\theta})$, os parâmetros $\boldsymbol{\theta}$ são ortogonais, isto é, os valores esperados das derivadas cruzadas da função de log-verossimilhança são iguais à zero. Neste caso é mais adequado o uso do algoritmo RS (Rigby e Stasinopoulos), considerado mais simples e que não utiliza o valor esperado das

derivadas cruzadas, o algoritmo RS é uma generalização do algoritmo utilizado por Rigby e Stasinopoulos (1996) no ajuste da média e dispersão de modelos aditivos.

Os algoritmos têm como objetivo maximizar a função de verossimilhança penalizada, dada pela Equação (2.7) no qual os hiperparâmetros $\boldsymbol{\lambda}$ são fixos. Nos modelos que são completamente paramétricos, os algoritmos maximizam a função de log-verossimilhança ℓ. Mais informações sobre os algoritmos CG e RS em Rigby e Stasinopoulos (2005).

2.1.5 Distribuições

Para a função de densidade de probabilidade $f(y|\boldsymbol{\theta})$ no modelo (2.1) utiliza-se qualquer família de distribuições sem que haja a obrigatoriedade de uma forma explícita para distribuição condicional da variável resposta y. Para a implementação dos modelos GAMLSS no R, a única restrição é que a função $f(y|\boldsymbol{\theta})$ e suas primeiras derivadas com relação aos parâmetros de $\boldsymbol{\theta}$ sejam calculáveis. Apesar de que as derivadas explícitas sejam preferíveis, também podem ser obtidas e utilizadas as derivadas numéricas. Nos modelos GAMLSS, é comum designar à variável resposta distribuições de probabilidade, disponíveis na função `gamlss`, a qual abrange uma ampla variedade de distribuições, indo além da família exponencial. A classe que engloba todas essas distribuições é definida como `gamlss.family`.

A notação utilizada para um GAMLSS é expressa por

$$y \sim D\left\{g_1(\theta_1) = t_1, g_2(\theta_2) = t_2, ..., g_p(\theta_p) = t_p\right\},$$

em que D é a distribuição da variável resposta, $\theta_1, ..., \theta_p$, são parâmetros de D; $g_1, ..., g_p$ são as funções de ligação e, $t_1, ..., t_p$ são as fórmulas dos modelos para os termos explanatórios e efeitos aleatórios nos preditores $\boldsymbol{\eta}_1, ..., \boldsymbol{\eta}_p$, respectivamente.

Nas Tabela 2 e 3 estão descritas algumas distribuições contínuas e discretas, disponíveis na biblioteca **`gamlss`** do *software R*, considerando suas respectivas nomenclaturas e funções de ligação padrão dos parâmetros. É importante destacar que o pacote `"gamlss"` também possibilita o ajuste de distribuições truncadas, censuradas e de misturas finitas (Rigby et al., 2019).

Tabela 2 – Distribuições contínuas implementadas no pacote `gamlss`.

Distribuição	**Nomenclatura**	**Função de ligação de parâmetro**			
		μ	σ	ν	τ
Exponencial Gaussiana	exGAUS()	identidade	log	log	-
Box-Cox Cole Green	BCCG()	identidade	log	identidade	-
Box-Cox Cole Green orig.	BCCGo()	log	log	identidade	-
Exponencial Potência de Box-Cox	BCPE()	identidade	log	identidade	log
Exponencial Potência orig.	BCPEo()	log	log	identidade	log
Box-Cox t	BCT()	identidade	log	identidade	log
Box-Cox t orig.	BCTo()	log	log	identidade	log
Beta	BE()	logito	logito	-	-
Exponencial	EXP()	log	-	-	-
Gamma	GA()	log	log	-	-
Beta Generalizada tipo 2	GB2()	log	log	log	log
Gamma Generalizada	GG()	log	log	identidade	-
Gaussiana Inversa Generalizada	GIG()	log	log	identidade	-
Gamma Inversa	IGAMMA()	log	log	-	-
Gaussiana Inversa	IG()	log	log	-	-
Log-Normal	LOGNO()	identidade	log	-	-
Log-Normal 2	LOGNO2()	log	log	-	-
Pareto 2	PARETO2()	log	log	-	-
Pareto 2 original	PARETO2o()	log	log	-	-
Weibull	WEI()	log	log	-	-
Weibull (PH)	WEI2()	log	log	-	-
Weibull (μ the mean)	WEI3()	log	log	-	-

Fonte: Adaptado de Rigby et al. (2019).

Tabela 3 – Distribuições discretas implementadas no pacote `gamlss`.

Distribuição	**Nomenclatura**	**Função de ligação de parâmetro**			
		μ	σ	ν	τ
Binomial	BI()	logito	-	-	-
Beta Binomial	BB()	logito	log	-	-
Poisson Generalizada	GPO()	log	log	-	-
Binomial Negativa Tipo 1	NBI()	log	log	-	-
Binomial Negativa Tipo 2	NBII()	log	log	-	-
Poisson	PO()	log	-	-	-
Poisson Inversa Gaussiana	PIG()	log	log	-	-
Sichel	SI()	log	log	identidade	-
Sichel (μ the mean)	SICHEL()	log	log	identidade	-
Poisson Inflada de Zeros	ZIP()	log	logito	-	-
Logarítmico	LG()	logito	-	-	-

Fonte: Adaptado de Rigby et al. (2019).

2.1.6 Seleção dos modelos

A seleção de modelos no GAMLSS inclui a escolha da melhor distribuição para a variável resposta, dos preditores adequados para os parâmetros da distribuição selecionada, das funções de ligação e dos hiperparâmetros. O critério de informação de Akaike generalizado (GAIC) pode ser utilizado para penalizar os modelos mais complexos e evitar sobreajustes, levando em consideração o número de parâmetros e de graus de liberdade que são usados no modelo o (Paiva; Freire; Cecatti, 2008). O critério de informação de Akaike (AIC) (Akaike, 1974) e o critério de informação Bayesiano (Schwarz, 1978) são casos especiais do critério GAIC (k) um caso se $k = 2$ recai no (AIC) e um outro caso $k = \ln(n)$ recai no (BIC). Os critérios penalizam os modelos mais complexos, aqueles com maiores números de parâmetros (Voudouris et al., 2012).

O GAIC é definido como

$$GAIC(k) = -2l(\hat{\boldsymbol{\theta}}) + (k \times df),$$

onde $l(\hat{\boldsymbol{\theta}})$ é o logaritmo da função verossimilhança ajustada e df são os graus efetivos de liberdade do modelo ajustado, k é constante e também é a penalidade para cada grau de liberdade utilizado. Refere-se a $-2l(\hat{\boldsymbol{\theta}})$, como desvio global, uma vez que o GAIC (k) é a estatística obtida pela adição do desvio global.

O desvio global (GDEV) é uma medida importante para a seleção de modelos em GAMLSS, dada por

$$GDEV = -2l(\hat{\boldsymbol{\theta}}),$$

em que, $l(\hat{\boldsymbol{\theta}})$ é o logaritmo da função de verossimilhança ajustada, apresentado na Equação (2.7).

Para a seleção do melhor modelo é usado o critério GAIC (k), ou seja, o modelo que apresentar menor valor de GAIC (k) para algum k escolhido, será indicado como o melhor ajuste, uma vez que o GAIC (k) penaliza modelos com muitos parâmetros.

2.1.7 Análise de resíduos

O uso de diagnóstico é um passo importante na modelagem estatística para a verificação do modelo e também na seleção do modelo. Para realizar o diagnóstico da qualidade do ajuste de um modelo existem diversas ferramentas estatísticas, entre elas, a análise dos resíduos que permite averiguar se as suposições do modelo foram satisfeitas (Stasinopoulos et al., 2017).

Nos modelos GAMLSS utiliza-se os resíduos dos quantis aleatórios normalizados, que podem ser usados para checar a distribuição utilizada para o ajuste do modelo, introduzidos por Dunn e Smyth (1996), e definido por,

$$r_i = \Phi^{-1}\left\{F(y_i;\hat{\theta})\right\},$$

onde Φ^{-1} é a inversa da função de distribuição acumulada de uma normal padrão, $F(.)$ é a função de distribuição acumulada adequada aos dados e $\hat{\boldsymbol{\theta}}$ é o vetor de parâmetros estimados. Note que, um modelo adequado tem os resíduos r_i seguindo a distribuição normal padrão.

Com o uso da função *resid()* do *software* R é possível obter os resíduos do modelo ajustado. A análise dos resíduos pode ser feita através de métodos gráficos, utilizando as funções *plot*() ou *wp*(). A função *plot*() é utilizada para a verificação geral dos resíduos, essa função produz quatro gráficos de resíduos quantílicos normalizados e um sumário de medidas resumo da distribuição dos resíduos. Já a função *wp*() retorna um gráfico *worm plot* único ou múltiplo para os modelos ajustados.

Um dos diferenciais na etapa do diagnóstico do GAMLSS, são os gráficos *worm plots* que foram introduzidos por Buuren e Fredriks (2001). Essa é uma ferramenta importante que pode ser aplicada para o uso de análise dos resíduos, que permite identificar inadequações no modelo globalmente ou dentro de um intervalo de uma variável explicativa. O *worm plot* é um gráfico normal quantil-quantil $(Q-Q)$ sem tendência para verificar alterações locais no domínio de uma dada covariável.

Os gráficos *worm plots*, podem ser utilizados para identificar regiões onde o modelo não é bem ajustado aos dados em estudo. O eixo vertical do gráfico worm plot, apresenta para cada observação a diferença entre a sua localização nas distribuições teórica e empírica. Quando os pontos são observados em conjuntos, formam uma curva que se assemelha a uma minhoca, mostrando como os dados se distanciam da distribuição assumida. Sendo assim, se os pontos estiverem situados entre as curvas elípticas, que representam o intervalo de confiança de 95%, o ajuste do modelo é considerado adequado (Stasinopoulos et al., 2017).

3 METODOLOGIA

Neste capítulo, será abordado o banco de dados empregado neste estudo, fornecendo detalhes sobre as variáveis que o compõem, além de informações relevantes sobre a fonte dos dados. Em seguida, serão apresentados conceitos e características relevantes sobre a distribuição inversa Gaussiana, uma componente essencial da família GAMLSS.

3.1 O banco de dados

Os dados utilizados no presente estudo foram obtidos por meio do site do Atlas do Desenvolvimento Humano no Brasil. Essa plataforma se destaca como uma ferramenta para obter informações estatísticas acerca do desenvolvimento humano no país. O Atlas reúne uma ampla gama de indicadores relacionados a diferentes aspectos da vida social, econômica e ambiental (IPEA; PNUD; FJP, 2022).

Para tanto, os dados utilizados são referentes a taxa de mortalidade por suicídio registrados nos estados brasileiros durante os anos de 2013 a 2017, considerando os 26 estados da federação e o Distrito Federal. Os estudos conduzidos por Santos e Barbosa (2017) e Santos, Barbosa e Severo (2020) abordaram dados de natureza semelhante aos dados em questão. O conjunto de dados analisado é referente a uma amostra com 135 observações. Foram utilizadas seis variáveis explicativas, sendo elas voltadas aos aspectos socioeconômicos, conforme a disponibilização dos dados públicos no recorte temporal utilizado. A taxa de mortalidade por suicídio foi a variável resposta e as demais variáveis descritas a seguir são as variáveis explicativas.

i) **Taxa de mortalidade por suicídio (TS)**: É o número de óbitos por suicídio registrados por 100.000 habitantes. Essa variável representa a taxa de mortalidade específica por suicídio em cada estado;

ii) **Renda per capita (RPC)**: É a renda média mensal das pessoas residentes em domicílios particulares permanentes. Essa variável mede o nível de renda das pessoas em cada estado;

iii) **Percentual de vulneráveis à pobreza (VP)**: Representa a proporção de indivíduos com renda domiciliar per capita igual ou inferior a R$255,00 mensais, em reais. Essa variável indica a porcentagem da população em situação de vulnerabilidade econômica;

iv) **Índice de Gini (GINI)**: Mede o grau de desigualdade econômica existente na distribuição de indivíduos segundo a renda domiciliar per capita. Um valor de 0 indica ausência de desigualdade, enquanto um valor próximo de 1 indica alta desigualdade econômica;

v) **Índice de Theil-L (THEIL)**: Mede a desigualdade na distribuição de indivíduos segundo a renda domiciliar per capita. É nulo quando não há desigualdade e tende a infinito quando a desigualdade é máxima. Essa medida exclui os indivíduos com renda domiciliar per capita nula;

vi) **Taxa de envelhecimento (TE)**: É a proporção da população idosa (65 anos ou mais de idade) em relação à população total. Essa variável reflete a proporção de idosos na população de cada estado;

vii) **Taxa de analfabetismo - 18 anos ou mais de idade (TA)**: Representa o percentual de pessoas com 18 anos ou mais que não sabem ler nem escrever um bilhete simples. Essa variável indica o nível de alfabetização da população adulta em cada estado.

3.2 Distribuição inversa Gaussiana

A distribuição inversa Gaussiana é composta por dois parâmetros de distribuições de probabilidade contínua com suporte em $(0, \infty)$. As distribuições de dois parâmetros são capazes de modelar somente localização e escala. A assimetria e a curtose dessas distribuições são definidas pelos valores de localização e escala. Ela tem uma maior assimetria positiva, sendo assim adequada para dados altamente distorcidos positivamente (Rigby et al., 2019).

Supondo que uma variável aleatória Y tem uma distribuição inversa Gaussiana com parâmetros μ e σ, e é denotada por IG (μ, σ), sua função densidade de probabilidade (p.d.f) é dada por

$$f_Y(y \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2 y^3}} exp\left[-\frac{1}{2\mu^2\sigma^2 y}(y-\mu)^2\right], \tag{3.1}$$

em que, $y > 0$, $\mu > 0$ e $\sigma > 0$, sendo μ a média da distribuição e σ o parâmetro de escala.

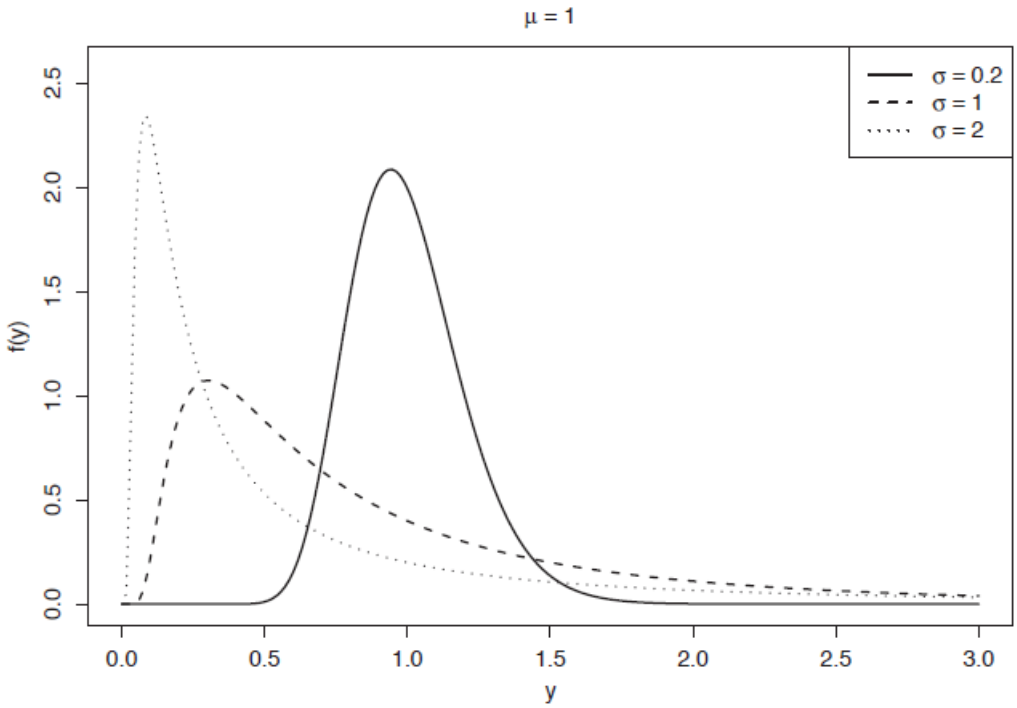
Foi proposta por Johnson, Kotz e Balakrishnan (1994) uma reparametrização, obtida ajustando $\sigma^2 = \frac{1}{\lambda}$. Logo a média e variância de Y são dadas por

$$E(Y) = \mu,$$

$$Var(Y) = \sigma^2\mu^3.$$

Na Figura 1 observamos o comportamento da função densidade de probabilidade para determinados valores de σ e para um valor de $\mu = 1$. Percebe-se que quando $\sigma \to 0$, a distribuição Gaussiana inversa fica mais simétrica em torno da média. De acordo com Paula (2004), Y se aproxima assintoticamente de uma distribuição normal com média μ e variância $\mu^3\sigma$. Sendo assim, a inversa Gaussiana se torna atrativa tanto para dados assimétricos quanto para dados simétricos que depende da forma cúbica da média.

Figura 1 – A distribuição inversa Gaussiana, $IG(\mu;\sigma)$, com $\mu = 1$ e $\sigma = 0.2, 1, 2$.



Fonte: Rigby et al. (2019).

Para modelagem da taxa de mortalidade por suicídio foi utilizado o software R (versão 4.2.2.) com o uso das bibliotecas **`gamlss`** e **`gamlss.dist`**. Como critério de escolha do melhor modelo, aplicou-se o critério de informação de Akaike generalizado (GAIC) (Paiva; Freire; Cecatti, 2008) com $k = 2$, os gráficos de diagnósticos dos resíduos e o gráfico *worm plot* (Buuren; Fredriks, 2001). Utilizou-se o teste de Shapiro-Wilk (Shapiro; Wilk, 1965) para verificar se os resíduos do modelo aproximam-se de uma amostra aleatória da distribuição normal padrão.

4 RESULTADOS E DISCUSSÃO

Neste capítulo, serão apresentados e discutidos os resultados obtidos utilizando a metodologia adotada. Inicialmente, foi realizada uma análise descritiva do conjunto de dados utilizado no estudo, que nos permitiu obter uma visão geral das características e distribuição dos dados. A análise descritiva incluiu cálculos estatísticos como média, mediana, desvio padrão, valores mínimos e máximos, assimetria e curtose para cada variável. Além disso, foi gerado o histograma para ilustrar as características da variável resposta. Após a análise descritiva, procedemos com o ajuste de GAMLSS ao conjunto de dados para analisar as relações entre as variáveis independentes e a variável resposta, explorando diferentes distribuições adequadas ao conjunto de dados.

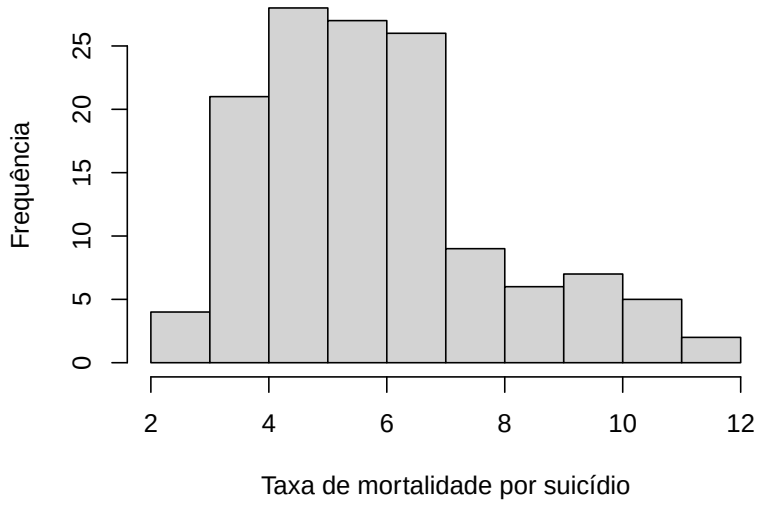
Tabela 4 – Estatísticas descritivas.

Estatística	**Variável**						
	TS	**RPC**	**VP**	**GINI**	**THEIL**	**TE**	**TA**
Mínimo	2,56	350,41	7,60	0,41	0,28	3,06	2,63
Máximo	11,91	1568,87	53,59	0,59	0,66	11,63	21,20
Média	5,82	686,73	30,63	0,51	0,47	7,31	9,90
Mediana	5,54	583,8	34,06	0,52	0,47	7,52	7,45
Desvio Padrão	2,07	248,8	13,77	0,04	0,08	1,83	5,6
Assimetria	0,8	1,45	-0,11	-0,25	0	-0,11	0,4
Curtose	0,12	2,46	-1,51	-0,38	-0,54	-0,41	-1,23

Fonte: Elaborado pela autora.

A partir das estatísticas descritivas apresentadas na Tabela 4, observa-se que a taxa de mortalidade por suicídio (TS) possui uma média de $5,82$ para todos os estados brasileiros, com um desvio padrão de $2,07$. Isso nos dá uma ideia da tendência central e da dispersão dos dados. Além disso, ao analisar a assimetria e a curtose, percebe-se que a variável TS é assimétrica positiva. Isso significa que a distribuição dos dados é inclinada para a direita, ou seja, há uma concentração maior de valores menores e alguns valores extremamente altos. Essa observação reforça a ideia de que os dados não seguem uma distribuição normal. Na Figura 2, é apresentado o histograma da variável resposta TS. Através desse gráfico, podemos visualizar a distribuição dos valores da variável e confirmar a assimetria positiva mencionada anteriormente. O histograma revela que a maioria dos valores estão concentrados na faixa inferior, com uma cauda longa estendendo-se para os valores mais altos. Essas informações são importantes para compreender a natureza e a distribuição dos dados da variável TS. A assimetria positiva sugere que existem fatores que influenciam o aumento da taxa de mortalidade por suicídio em determinados estados brasileiros, resultando em valores mais altos do que a média. Essa análise exploratória nos permite ter uma visão inicial das características da variável resposta.

Figura 2 – Histograma da taxa de mortalidade por suicídio.



Fonte: Elaborado pela autora.

Ao analisar as covariáveis que compõem o presente estudo, observa-se algumas características importantes. Os valores apresentados na Tabela 4 revelam discrepâncias significativas entre os valores mínimos e máximos de algumas variáveis, o que pode estar relacionado às diferenças populacionais e territoriais entre os estados brasileiros observados neste trabalho. A variável Renda per capita (RPC) possui um valor médio de $R\$686,73$. No entanto, há uma variação considerável entre o valor mínimo de $R\$350,41$ e o valor máximo de $R\$1568,87$. O desvio padrão de $248,8$ indica uma dispersão significativa dos dados em relação à média. Isso sugere que existem diferenças consideráveis na renda per capita entre os estados brasileiros.

Quanto ao Percentual de vulneráveis à pobreza (VP), observa-se que $30,63\%$ da população dos estados brasileiros se encontra nessa situação. Essa informação é relevante para entender a proporção de pessoas em situação de vulnerabilidade econômica nos diferentes estados. A variável Índice de Gini (GINI), que mede a desigualdade de renda, apresenta um valor médio de $0,51$, com um desvio padrão de $0,04$. Isso indica que há uma certa variação na desigualdade de renda entre os estados. Já o Índice de Theil-L (THEIL), que também mede a desigualdade, possui uma média de $0,47$ e um desvio padrão de $0,08$. Esses valores indicam que existe uma certa dispersão dos dados em relação à média, sugerindo variações na desigualdade entre os estados brasileiros. A Taxa de envelhecimento (TE) apresenta uma média de $7,31$ e um desvio padrão de $1,83$. Esses valores indicam uma certa variação na proporção de idosos em relação à população total nos diferentes estados. Por fim, a Taxa de analfabetismo (TA) possui uma média de $9,90$ para os 27

estados brasileiros, com um desvio padrão de $5,6$. Isso indica que há variações consideráveis na taxa de analfabetismo entre os estados. Essas informações sobre as covariáveis são relevantes para compreender as características socioeconômicas e demográficas dos estados brasileiros. Elas fornecem informações iniciais sobre as diferenças entre os estados e podem ser levadas em consideração durante a análise dos resultados e interpretação dos modelos.

Em seguida, foi verificado o quanto as variáveis explicativas estão relacionadas com a variável resposta. Considerando que os dados apresentam desvio de normalidade em sua estrutura, utilizou-se o coeficiente de correlação de Spearman (ρ) e seus respectivos p-valores para verificar a significância destas relações.

De acordo com os coeficientes de correlação de Spearman observados na Tabela 5, a relação entre a variável resposta TS e a covariável RPC apresenta correlação positiva fraca e significativa (p-valor$< 0,05$). Já a relação entre a variável TS e as covariáveis VP, GINI, THEIL e TA apresentam correlação negativa fraca, porém significativas (p-valor$< 0,05$). A variável TS e a covariável TE tem correlação bem fraca com p-valor$> 0,05$, indicando que a correlação não é estatisticamente significativa. Conforme demonstrado no estudo de Schnitman et al. (2010), através da análise de correlação de Spearman, também verificou-se correlação negativa com significância estatística entre os indicadores de desigualdade social (índice de Gini e índice de Theil-L) e a taxa de mortalidade por suicídio.

Tabela 5 – Teste de correlação de Spearman.

Variáveis	**Correlação de Spearman**	**ρ-valor**
(TS, RPC)	0,29	0,0007
(TS, VP)	-0,35	0,0000
(TS, GINI)	-0,30	0,0004
(TS, THEIL)	-0,30	0,0003
(TS, TE)	0,09	0,3182
(TS, TA)	-0,22	0,0089

Fonte: Elaborado pela autora.

No próximo passo, podem ser utilizas essas informações para realizar análises mais aprofundadas e aplicar a metodologia proposta, como o ajuste dos GAMLSS mencionados anteriormente, a fim de explorar as relações entre as variáveis explicativas e a variável resposta, considerando também as distribuições e estruturas de dispersão adequadas aos dados.

No contexto de GAMLSS, existe uma variedade de distribuições disponíveis para modelar a variável resposta. Foi utilizado o pacote **`gamlss`** e a função **`fitDist()`** para estimar a distribuição da variável resposta "Taxa de mortalidade por suicídio". A função **`fitDist()`** ajustará diferentes distribuições aos dados de TS e, a partir disso, é possível avaliar qual distribuição se ajusta melhor aos dados. O argumento `type` é usado para especificar o tipo de distribuição a ser ajustada, ao selecionar a opção "`realplus`", está

sendo especificando o interesse em distribuições contínuas com suporte no intervalo $(0, \infty)$, ou seja, apenas valores positivos.

Na Tabela 6, são listadas diferentes distribuições que podem ser ajustadas à variável resposta, juntamente com os valores correspondentes do GAIC. Os valores de GAIC são dispostos em ordem crescente, refletindo a qualidade do ajuste de cada distribuição aos dados observados. É importante ressaltar que quanto menor o valor de GAIC, melhor é o ajuste da distribuição aos dados. As distribuições na tabela são denotadas pelas siglas utilizadas no pacote **`gamlss`**. Para obter informações detalhadas sobre o significado e as propriedades estatísticas de cada distribuição, recomenda-se consultar Stasinopoulos et al. (2017).

Tabela 6 – Principais distribuições dos modelos probabilísticos ajustados via pacote `gamlss` utilizando o Critério de Informação de Akaike Generalizado.

Distribuição	**GAIC**	**Distribuição**	**GAIC**
IG	559,4516	BCT	564,0457
LOGNO2	560,1567	BCTo	564,0457
LOGNO	560,1567	GB2	564,1586
IGAMMA	561,1056	exGAUS	565,1510
GIG	561,4009	WEI	578,3606
GG	562,0363	WEI2	578,3606
BCCG	562,0457	WEI3	578,3606
BCCGo	562,0457	EXP	747,6610
BCPEo	562,4703	PARETO2o	749,6610
BCPE	562,4703	PARETO2	749,6611
GA	563,0546	GP	749,6621

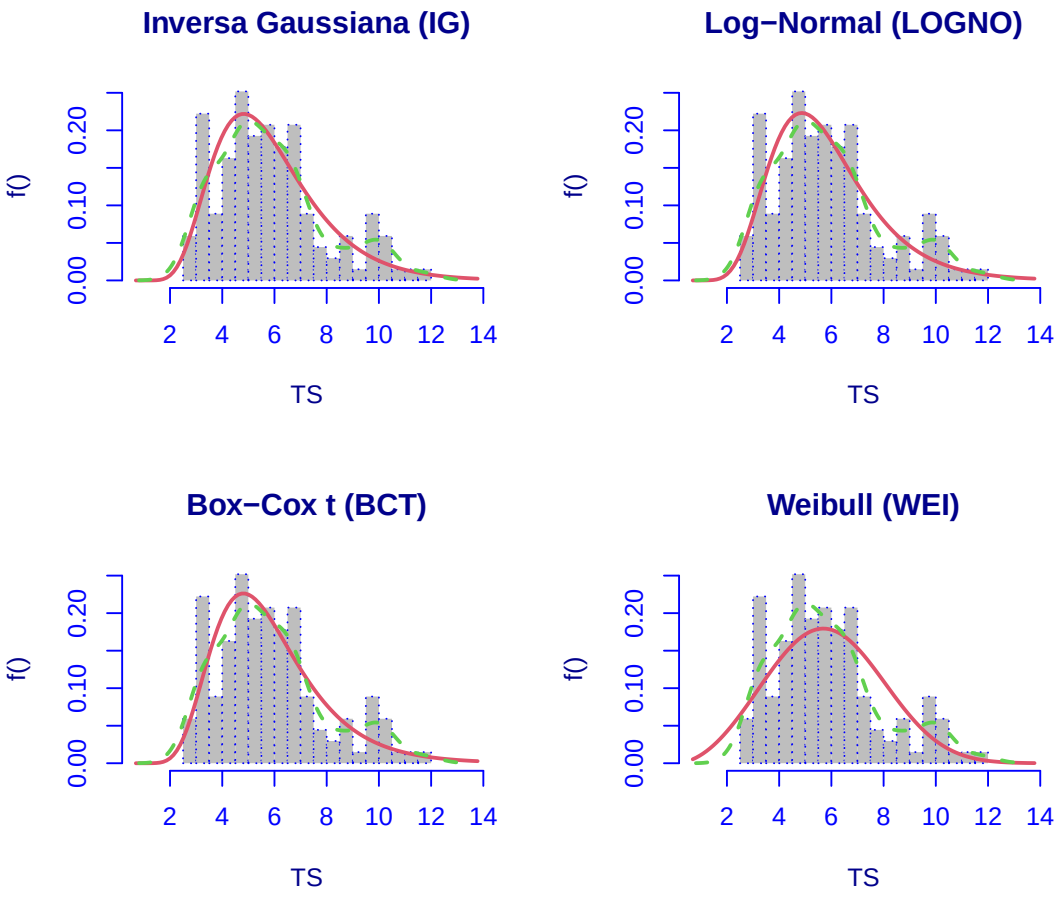
Fonte: Elaborado pela autora.

Entre as distribuições apresentadas na Tabela 6, destaca-se a distribuição inversa Gaussiana (IG), que obteve o menor valor de GAIC. Esse resultado indica que a distribuição IG apresenta o melhor ajuste aos dados observados em comparação com as outras distribuições listadas. Além disso, outras distribuições, como a log normal, Box-Cox t e Weibull, também obtiveram valores de GAIC que as colocam em posições intermediárias na tabela. Embora não sejam as distribuições com os melhores ajustes, ainda são consideradas como opções razoáveis para ajustar a variável resposta. Em um estudo conduzido por Barajas e Naranjo (2022), observou-se que, inicialmente, a distribuição inversa Gaussiana destacou-se ao exibir o menor valor de GAIC durante o processo de seleção das distribuições que proporcionaram o melhor ajuste à variável resposta.

Para ajustar os dados, foram consideradas as distribuições inversa Gaussiana, log-normal, Box-Cox t e Weibull. Com o intuito de analisar a adequação dessas distribuições aos dados, foram construídos histogramas utilizando a função **`histDist()`**. Na Figura 3, pode-se observar os resultados desses histogramas, neles estão respresentados os valores

observados da taxa de mortalidade por suicídios. A linha vermelha representa a função densidade paramétrica e a linha azul representa a densidade estimada não-parametricamente.

Figura 3 – Ajuste das distribuições IG, LOGNO, BCT e WEI à variável resposta.



Fonte: Elaborado pelo autora.

Essa visualização permite comparar a distribuição teórica ajustada com a distribuição observada dos dados. Uma boa adequação entre a distribuição ajustada e os dados é indicada quando a função densidade paramétrica se aproxima da densidade estimada não-paramétrica. Ao observar a Figura 3, nota-se que as distribuições inversa Gaussiana, log-normal e Box-Cox t se ajustam melhor.

Para o ajuste dos modelos apresentados a seguir, foram consideradas todas as seis variáveis explicativas: RPC, VP, GINI, TA, THEIL e TE. Inicialmente foi utilizada a função **`dropterm`** que permite adicionar ou remover variáveis explicativas em um modelo. Essa função foi empregada para avaliar a contribuição de cada variável no ajuste do GAMLSS para cada distribuição considerada. Esse procedimento possibilitou uma seleção adequada das variáveis explicativas, visando encontrar o melhor ajuste para cada distribuição em análise.

A função **`stepGAIC.VR`** foi utilizada para construir modelos individualmente para cada parâmetro das distribuições. De acordo com Stasinopoulos et al. (2017), é recomendado respeitar a hierarquia dos parâmetros. Portanto, o ajuste foi iniciado para o parâmetro μ e, em seguida, para o parâmetro σ. Ao utilizar a função **`stepGAIC.VR`**, foi possível realizar a seleção automática de variáveis e obter modelos otimizados para cada parâmetro das distribuições consideradas. Essa abordagem ajuda a identificar quais variáveis são mais relevantes para cada parâmetro.

Os resultados obtidos com o uso das funções mencionadas anteriormente são apresentados na Tabela 7. Ao comparar os diferentes modelos, observa-se que a distribuição inversa Gaussiana apresenta os melhores valores para todos os critérios avaliados que estão apresentados na tabela. Esses resultados indicam que a distribuição inversa Gaussiana é a mais adequada para descrever os dados em questão em relação às outras distribuições consideradas.

Tabela 7 – Critérios de seleção para os modelos ajustados.

Distribuição	**Desvio global**	AIC
IG	504,5281	524,5281
LOGNO	518,086	532,086
BCT	516,2795	534,2795
WEI	524,5965	544,5965

Fonte: Elaborado pela autora.

Na Tabela 8, estão listados os parâmetros estimados para cada distribuição, juntamente com as variáveis explicativas incluídas em cada modelo. Essa tabela fornece uma visão geral dos resultados obtidos após a aplicação das técnicas de seleção de variáveis mencionadas anteriormente.

Tabela 8 – Variáveis incluídas nos preditores dos parâmetros dos modelos com o uso da função `stepGAIC.VR`.

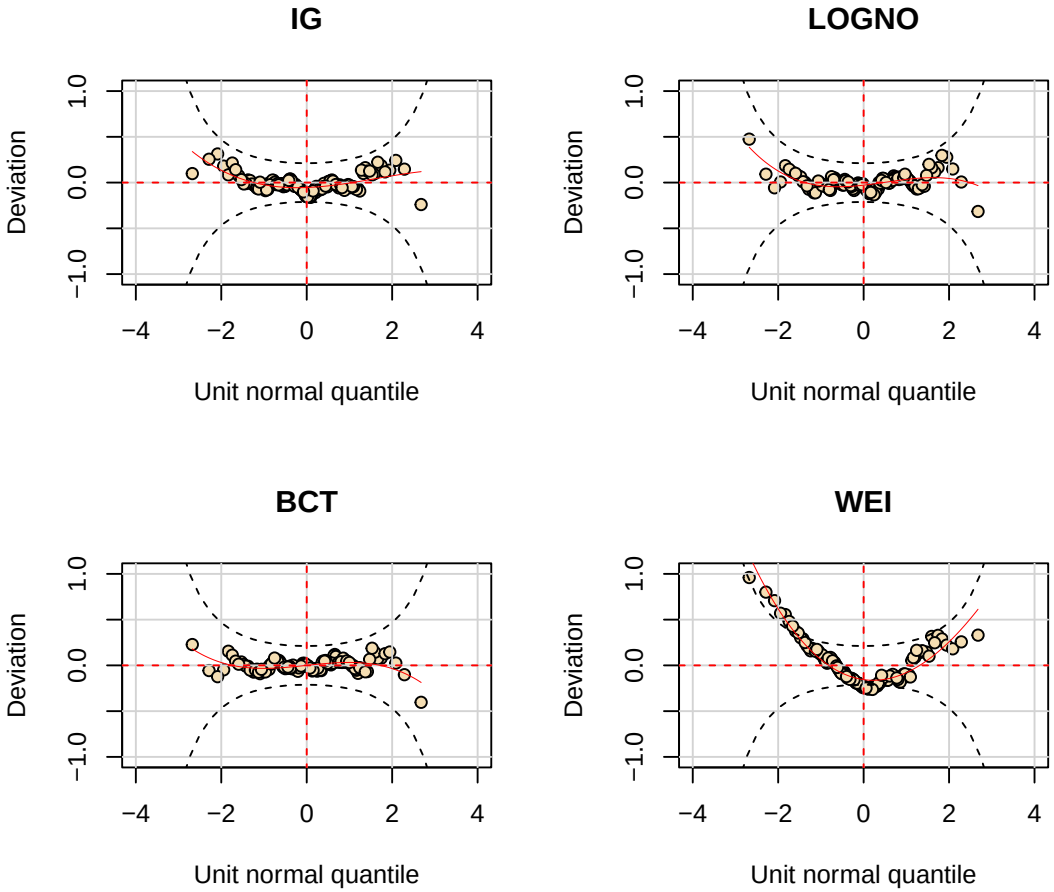
Distribuições	**Preditores**	**RPC**	**VP**	**GINI**	**TA**	**THEIL**	**TE**
IG	μ	X	X	X		X	X
	σ	X		X	X		
LOGNO	μ	X	X	X		X	X
	σ						
BCT	μ	X	X	X		X	X
	σ						
WEI	μ	X	X	X		X	
	σ	X	X	X			X

Fonte: Elaborado pelo autora.

Na Figura 4 são apresentados os gráficos *worm plot* de cada modelo ajustado. Os *worm plots*, também conhecidos como gráficos de minhocas, são utilizados para identificar

inadequações no modelo. Ao analisar os *worm plots* dos modelos, observa-se que tanto o modelo IG quanto o modelo BCT apresentaram bons resultados. Todos os pontos estão dentro da banda de confiança de 95% e a maioria deles estão concentrados próxima à reta $y = 0$, o que indica bons ajustes.

Figura 4 – Gráficos *Worm plot* dos modelos ajustados.



Fonte: Elaborado pela autora.

No entanto, é importante resaltar que a distribuição inversa Gaussiana se destaca com os melhores/menores resultados nas métricas de desvio global e critério de GAIC. Esse modelo obteve um desempenho superior nas métricas avaliadas, justificando assim a escolha da distribuição inversa Gaussiana como a candidata mais adequada para o ajuste dos dados em estudo. Conforme demonstrado no estudo de Morita (2017), utilizando a análise de confiabilidade, a distribuição inversa Gaussiana também apresentou os melhores resultados.

O processo prosseguiu com a aplicação da função **`stepGAIC.CH`** para a identificação de um modelo apropriado, utilizando um spline de suavização cúbica como técnica de suavização para o modelo IG. O procedimento incluiu a construção de modelos individua-

lizados para cada parâmetro da distribuição da variável de resposta. Inicialmente, o foco foi no parâmetro μ, e posteriormente, estendeu-se para modelar o parâmetro σ.

Uma vez que as variáveis independentes foram selecionadas por meio das funções mencionadas anteriormente, prosseguiu-se com a aplicação da função **`stepGAIC.CH`** para efetuar o ajuste do parâmetro μ. Os resultados obtidos por meio dessa abordagem indicam que o modelo mais adequado incorpora termos de suavização para as seguintes variáveis: Percentual de Vulneráveis à Pobreza (VP), Índice de Gini (GINI) e Taxa de Envelhecimento (TE). Isso sugere que essas variáveis podem ter relações complexas e não lineares com a variável dependente.

Logo após a obtenção de um modelo adequado para o parâmetro μ, prosseguiu-se com etapas semelhantes para o ajuste do parâmetro σ. Foi realizada uma análise da relevância estatística das variáveis previamente selecionadas. Após confirmar sua relevância, a função **`stepGAIC.CH`** foi novamente aplicada, recorrendo desta vez ao algoritmo RS para alcançar a convergência necessária do modelo. Entretanto, em contraste com o processo anterior, desta vez não foram identificados termos de suavização para as variáveis.

O GAMLSS gerado para todos os parâmetros da distribuição pode ser descrito pela seguinte expressão:

$$g_1(\mu) = \beta_{10} + \beta_{11}RPC + \beta_{12}cs(VP) + \beta_{13}cs(GINI) + \beta_{14}THEIL + \beta_{15}cs(TE), e \tag{4.1}$$

$$g_2(\sigma) = \beta_{20} + \beta_{21}RPC + \beta_{22}TA + \beta_{23}GINI. \tag{4.2}$$

Os resultados da análise utilizando o GAMLSS e as funções discutidas anteriormente, estão apresentados na Tabela 9. Nela, são destacados os principais parâmetros, incluindo as estimativas calculadas, os erros padrão associados a essas estimativas e os p valores correspondentes. Esses p valores são resultado do teste de significância dos coeficientes associados às covariáveis. Esse teste busca determinar se a presença dessas covariáveis exerce uma influência estatisticamente significativa sobre a variável resposta.

Tabela 9 – Estimativas dos parâmetros, erro padrão e p-valor para o modelo IG.

Parâmetro	**Variável**	**Estimativa**	**Erro Padrão**	***p*-valor**
μ	Intercepto	12,294	0,4553	0,0000
	RPC	-0,0007	0,0002	0,0016
	cs(VP)	-0,0223	0,0061	0,0004
	cs(GINI)	-32,913	1,3756	0,0000
	THEIL	16,337	0,8984	0,0000
	cs(TE)	-0,0215	0,0162	0,1852
σ	Intercepto	-1,7422	1,0573	0,1021
	RPC	-0,0025	0,0005	0,0000
	TA	-0,0648	0,0219	0,0037
	GINI	3,5975	2,2786	0,1171

Fonte: Elaborado pela autora.

De acordo com os resultados apresentados na Tabela 9, no que diz respeito ao parâmetro μ, ao nível de 5% de significância, é possível observar que as variáveis RPC, VP, GINI E THEIL apresentaram significância estatística com um p-valor inferior a $0,05$. Isso evidencia uma relação significativa com a taxa de mortalidade por suicídio. Por outro lado, à variável taxa de envelhecimento (TE) apresenta um p-valor superior a $0,05$, sugerindo então que não há evidências estatísticas suficientes para concluir que o coeficiente é diferente de zero, com base nesse resultado e considerando um nível de significância de 5%, não é possível afirmar que essa variável tem um efeito estatisticamente significativo no resultado do modelo. Essa mesma situação é observada para o termo intercepto e a variável Índice de Gini (GINI), quando consideramos o parâmetro σ, dado que os p-valores associados a esses elementos ultrapassam o limite de $0,05$. Por outro lado, as variáveis RPC e TA apresentaram significância estatística, com um p-valores inferiores a $0,05$, indicando uma relação significativa com a taxa de mortalidade por suicídio.

De acordo com as estimativas destacadas na Tabela 9, observa-se uma relação estatisticamente significativa e inversa entre a variável renda per capita (RPC) e a taxa de mortalidade por suicídio. Em outras palavras, os dados sugerem que níveis mais elevados de renda per capita estão associados a uma diminuição na taxa de suicídio. Por outro lado, o índice de Theil apresenta relação direta e significante com o suicídio, indicando que à medida que a desigualdade de renda aumenta, a taxa de suicídio também tende a crescer. Em um estudo desenvolvido por Amaral (2019), também foi observado relações estatisticamente significativas da renda per capita e do índice de Theil com a taxa de suicídio.

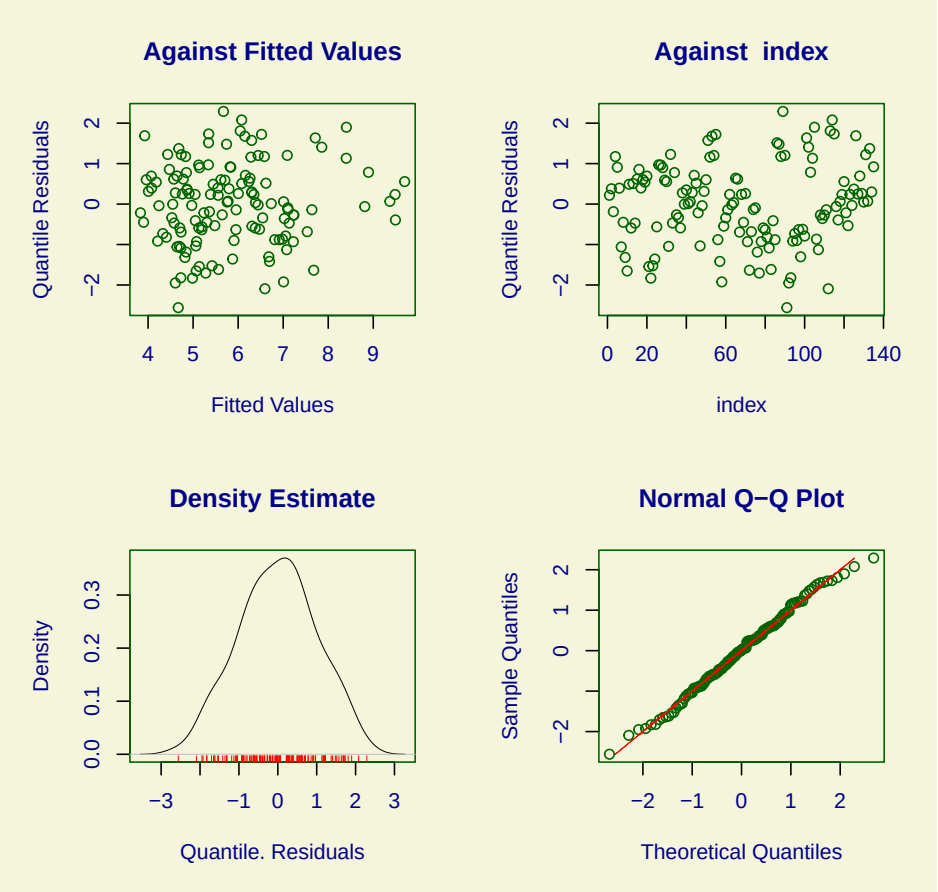
As variáveis percentual de vulneráveis à pobreza (VP) e índice de Gini (GINI) demonstraram uma relação estatisticamente significativa e inversa com a taxa de suicídio. Observou-se que um aumento no percentual de vulneráveis à pobreza está associado a uma redução na taxa de suicídio. Por outro lado, constatou-se que um aumento no índice de

Gini está relacionado a uma elevação na taxa de suicídio, sugerindo que maior desigualdade econômica pode estar associada a um aumento nos casos de suicídio.

No contexto do parâmetro σ, foi observada uma relação significativa e inversa entre a taxa de analfabetismo (TA) e a taxa de suicídio.

Para exemplificar, Schnitman et al. (2010) mostrou em seu estudo associação inversa entre o Índice de Gini e a taxa de suicídio. Os resultados de Ribeiro, Ferreira e Oliveira (2022), que analisaram a taxa de suicídio para o sexo masculino, também indicaram que o índice de Gini possui associação inversa com o suicídio. Estudos internacionais Taylor et al. (2004), Middleton, Sterne e Gunnell (2006), Kim et al. (2010), Chang et al. (2011) também constataram associação inversa entre condições socioeconômicas e suicídio.

Figura 5 – Gráficos residuais do modelo IG obtidos através da função plot().



Fonte: Elaborado pela autora.

Stasinopoulos et al. (2017) destacam a importância de avaliar a adequação dos resíduos gerados após o ajuste de um modelo GAMLSS. Para verificar a adequação do modelo, foram utilizadas as ferramentas gráficas obtidas com o uso da função plot(). Na Figura 5, são apresentados os gráficos utilizados para verificar os resíduos quantílicos normalizados do modelo ajustado. Observando os dois gráficos superiores dos resíduos, é notável a ausência de qualquer padrão para os resíduos. Eles exibem uma dispersão

aleatória em relação à linha horizontal de referência situada em 0. Por outro lado, o gráfico localizado no canto inferior esquerdo apresenta uma forma semelhante à da função densidade da normal $(0, 1)$. Além disso, no gráfico QQ-plot, pode-se observar que os resíduos estão próximos à reta diagonal $y = x$, indicando que eles estão bem alinhados com a distribuição normal. Dessa forma, a análise dos gráficos dos resíduos aponta que o modelo ajustado se adequa de forma satisfatória aos dados.

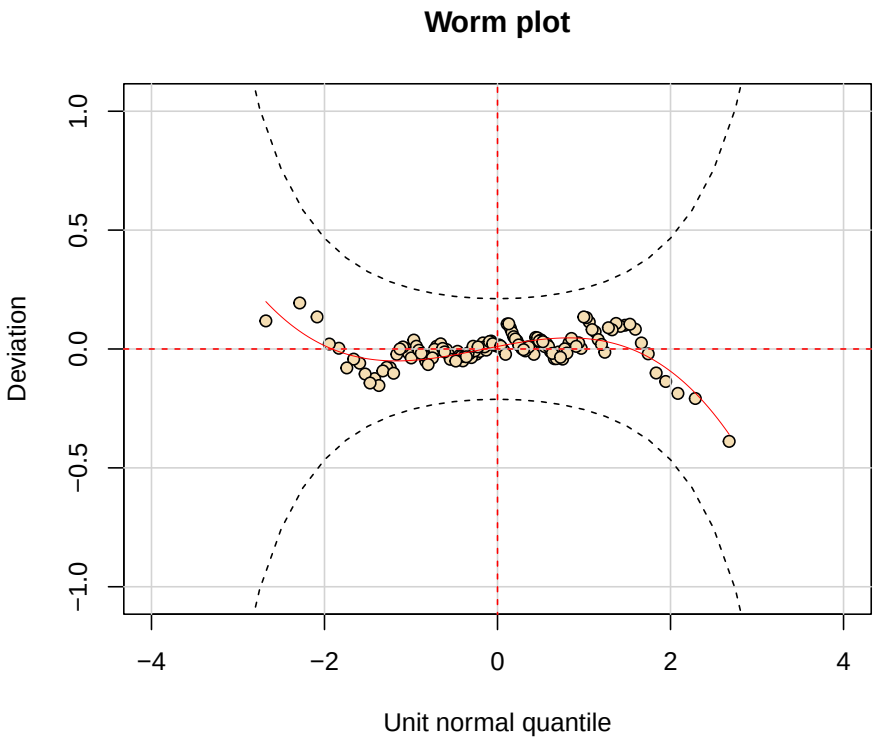
Tabela 10 – Medidas descritivas dos resíduos do modelo GAMLSS ajustado.

Parâmetro	**Valores estimados**
Média	-0,0018
Variância	1,0057
Coeficiente de assimetria	-0,0590
Coeficiente de curtose	2,4944
Coeficiente de correlação de Filliben	0,9975

Fonte: Elaborado pela autora.

Na Tabela 10, são apresentadas as medidas descritivas referentes aos resíduos do modelo ajustado, obtidas por meio da função plot(). Ao examinar a tabela, fica evidente que os resíduos têm uma média próxima a zero, com um valor de $-0,0018$. Quanto à variância, seu valor de 1,0057 se mostra relativamente próximo a 1, indicando que a dispersão dos resíduos é comparável a dos dados originais. A assimetria apresenta um valor igual a $-0,0590$ próximo de zero, que corresponde a uma curva platicúrtica, sugerindo que a distribuição dos resíduos é aproximadamente simétrica. A curtose é igual a $2,4944$, relativamente próxima de 3, indicando que a distribuição dos resíduos tem uma forma relativamente próxima de uma distribuição normal. O coeficiente de correlação de Filliben igual a $0,9975$ próximo de 1, sugerindo uma alta correlação entre os resíduos observados e os esperados. Também foi utilizado o teste de Shapiro-Wilk, para avaliar a hipótese de normalidade nos resíduos, considerando um nível de 5% de significância. O resultado apresentou um p-valor de $0,80$, indicando que não há evidências para rejeitar a hipótese nula de que os resíduos seguem uma distribuição normal. Portanto, as características observadas referentes as medidas descritivas e ao teste de Shapiro-Wilk, indicam que os resíduos seguem uma distribuição que se assemelha à normal padrão.

Figura 6 – `Worm plot` dos resíduos do modelo GAMLSS ajustado.



Fonte: Elaborado pela autora.

Na Figura 6, é exibido o gráfico de `worm plot` do modelo ajustado. Nesse gráfico, é perceptível que todos os pontos estão contidos dentro da faixa de confiança de 95%. Além desse aspecto, é possível notar que a grande maioria dos pontos encontram-se concentrados próximos a reta $y = 0$. Essa observação sugere que o modelo apresenta um ajuste satisfatório aos dados.

5 CONCLUSÃO

Ao longo deste trabalho, foram explorados os fundamentos teóricos da classe GAMLSS. Além de abordar aspectos cruciais de inferência e diagnóstico, destaca-se a flexibilidade dessa classe de modelos em ajustar uma ampla família de distribuições à variável resposta. É importante ressaltar também a capacidade dos GAMLSS em modelar todos os parâmetros da distribuição da variável resposta em função da variável explanatória. Essa classe oferece a possibilidade de incorporar termos suavizados no modelo, ampliando assim sua capacidade de capturar padrões complexos nos dados.

A aplicação do GAMLSS indicou que a distribuição inversa Gaussiana, com o menor valor de GAIC, proporcionou o melhor ajuste para a taxa de mortalidade por suicídio, permitindo a modelagem dos parâmetros de média e dispersão. A validação da qualidade do ajuste foi conduzida por meio de análises diagnósticas abrangentes, a análise gráfica dos resíduos, juntamente com medidas descritivas dos resíduos e o gráfico *worm-plot*, confirmando a adequação do modelo final. Esses resultados fortalecem a confiança na capacidade do modelo IG em capturar a variabilidade na taxa de mortalidade por suicídio. As estimativas do modelo final revelaram que variáveis como renda per capita, percentual de vulneráveis à pobreza, índice de Gini, índice de Theil e taxa de analfabetismo desempenham um papel significativo na influência da taxa de mortalidade por suicídio. Suas relações identificadas indicam a capacidade dessas variáveis em contribuir na taxa de mortalidade por suicídio.

REFERÊNCIAS

AGUIAR, C. R.; CARVALHO, M. de O. G. Lesões autoprovocadas e suicídios 2009 - 2018. Boletim epidemiológico 002/2019: Divisão de Vigilância de Doenças e Agravos Não Transmissíveis, Rio de Janeiro, ano 2019, ed. 002, p. 1-43. 2019. Disponível em: <http://www.riocomsaude.rj.gov.br/Publico/MostrarArquivo.aspx?C=4cM0P9oa57k%3D>. Acesso em: 29 jul. 2023.

AITKIN, Murray et al. **Statistical modelling in R**. Oxford: Oxford University Press, 2009

AKAIKE, Hirotugu. A new look at the statistical model identification. **IEEE transactions on automatic control**, v. 19, n. 6, p. 716-723, 1974.

AMARAL, Stefany Silva. Suicídio no RN e sua relação com determinantes espaciais, urbanização, desenvolvimento e outros fatores socioeconômicos. **Revista Brasileira de Estudos Regionais e Urbanos**, v. 13, n. 2, p. 288-308, 2019.

BARAJAS, Freddy Hernández; NARANJO, Yeison Yovany Ocampo. Modelos GAMLSS como una alternativa para mejorar el proceso de recubrimiento de instrumentos quirúrgicos con cromoduro: Modelos GAMLSS para estudiar procesos de recubrimiento. **Revista EIA**, v. 19, n. 38, p. 3818 pp. 1-14, 2022.

BRASIL, M. da S. Perfil epidemiológico dos casos notificados de violência autoprovocada e óbitos por suicídio entre jovens de 15 a 29 anos no Brasil, 2011 a 2018. **Bol Epidemiológico**, v. 50, 2019.

BUUREN, Stef van; FREDRIKS, Miranda. Worm plot: a simple diagnostic device for modelling growth reference curves. **Statistics in medicine**, v. 20, n. 8, p. 1259-1277, 2001.

CHANG, Shu-Sen et al. Geography of suicide in Taiwan: spatial patterning and socioeconomic correlates. **Health & place**, v. 17, n. 2, p. 641-650, 2011.

CLEVELAND, William S.; GROSSE, Eric; SHYU, William M. Local regression models. In: **Statistical models in S**. Routledge, 2017. p. 309-376.

COLE, Timothy J.; GREEN, Pamela J. Smoothing reference centile curves: the LMS method and penalized likelihood. **Statistics in medicine**, v. 11, n. 10, p. 1305–1319, 1992.

DUNN, Peter K.; SMYTH, Gordon K. Randomized quantile residuals. **Journal of Computational and graphical statistics**, v. 5, n. 3, p. 236-244, 1996.

DURKHEIM, Emile. O Suicídio: Estudo de Sociologia (Trad. M. Stahel). 2000.

GONÇALVES, Ludmilla RC; GONÇALVES, Eduardo; OLIVEIRA JÚNIOR, Lourival

Batista de. Determinantes espaciais e socioeconômicos do suicídio no Brasil: uma abordagem regional. **Nova Economia**, v. 21, p. 281-316, 2011.

HASTIE, T.; TIBSHIRANI, R. Generalized Additive Models. **Statistical Science**, Institute of Mathematical Statistics, v. 1, n. 3, p. 297 – 310, 1986. Disponível em: <https://doi.org/10.1214/ss/1177013604>.

HASTIE, T.; TIBSHIRANI, R. **Generalized Additive Models**. Taylor & Francis, 1990. (Chapman & Hall/CRC Monographs on Statistics & Applied Probability). ISBN 9780412343902. Disponível em: <https://books.google.com.br/books?id=qa29r1Ze1coC>.

IHAKA, Ross; GENTLEMAN, Robert. R: a language for data analysis and graphics. **Journal of computational and graphical statistics**, v. 5, n. 3, p. 299–314, 1996.

IPEA; PNUD; FJP. Atlas do Desenvolvimento Humano no Brasil. Disponível em:<http://www.atlasbrasil.org.br/consulta/planilha>. Acesso em: 17 mar. 2023.

JOHNSON, N.; KOTZ, S.; BALAKRISHNAN, N. **Continuous Univariate Distributions, Volume 1**. Wiley, 1994. (Wiley Series in Probability and Statistics). ISBN 9780471584957. Disponível em: <https://books.google.com.br/books?id=swkwzwEACAAJ>.

KIM, Myoung-Hee et al. Socioeconomic inequalities in suicidal ideation, parasuicides, and completed suicides in South Korea. **Social science & medicine**, v. 70, n. 8, p. 1254-1261, 2010.

LEE, Youngjo; NELDER, John A. Hierarchical generalized linear models. **Journal of the Royal Statistical Society Series B: Statistical Methodology**, v. 58, n. 4, p. 619-656, 1996.

LEITE, Rayane Santos. **Aplicação de modelos aditivos generalizados para locação, escala e forma em clones de Eucalyptus spp., no polo gesseiro do Araripe-PE**. 2019. Tese de Doutorado. Dissertação (Mestrado)–Universidade Federal Rural de Pernambuco, Programa de Pós-Graduação em Biometria e Estatística Aplicada, Recife, BR-PE.

MIDDLETON, Nicos; STERNE, Jonathan AC; GUNNELL, David. The geography of despair among 15–44-year-old men in England and Wales: putting suicide on the map. **Journal of Epidemiology & Community Health**, v. 60, n. 12, p. 1040-1047, 2006.

MORITA, Lia Hanna Martins. Degradation modeling for reliability analysis with time-dependent structure based on the inverse gaussian distribution. 2017.

NELDER, John Ashworth; WEDDERBURN, Robert WM. Generalized linear models. **Journal of the Royal Statistical Society Series A: Statistics in Society**, v. 135, n. 3, p. 370-384, 1972.

PAIVA, Cláudio Sérgio Medeiros; FREIRE, Djacyr Magna Cabral; CECATTI, José Guilherme. Modelos aditivos generalizados para posição, escala e forma (gamlss) na modelagem de curvas de referência. **Rev. bras. ciênc. saúde**, p. 289-310, 2008.

PAULA, Gilberto Alvarenga. **Modelos de regressão: com apoio computacional**. São Paulo: IME-USP, 2004.

PAWITAN, Yudi. **In all likelihood: statistical modelling and inference using likelihood**. Oxford University Press, 2001.

RIBEIRO, Anna Carolina Mendonça Lemos; FERREIRA, Pedro Cavalcanti Gonçalves; OLIVEIRA, João Maria de. Determinantes socioeconômicos do suicídio nos estados brasileiros: análise de dados em painel de 2010 a 2015. 2022.

RIGBY, Robert A.; STASINOPOULOS, D. M. A semi-parametric additive model for variance heterogeneity. **Statistics and Computing**, v. 6, p. 57-65, 1996.

RIGBY, Robert A.; STASINOPOULOS, D. Mikis. Generalized additive models for location, scale and shape. **Journal of the Royal Statistical Society Series C: Applied Statistics**, v. 54, n. 3, p. 507-554, 2005.

RIGBY, Robert A.; STASINOPOULOS, D. Mikis. Using the Box-Cox t distribution in GAMLSS to model skewness and kurtosis. **Statistical Modelling**, v. 6, n. 3, p. 209-229, 2006.

RIGBY, Robert A. et al. **Distributions for modeling location, scale, and shape: Using GAMLSS in R**. CRC press, 2019.

ROYSTON, Patrick; ALTMAN, Douglas G. Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling. **Journal of the Royal Statistical Society Series C: Applied Statistics**, v. 43, n. 3, p. 429-453, 1994.

SANTOS, Emelynne Gabrielly de Oliveira; BARBOSA, Isabelle Ribeiro. Conglomerados espaciais da mortalidade por suicídio no nordeste do Brasil e sua relação com indicadores socioeconômicos. **Cadernos Saúde Coletiva**, v. 25, p. 371-378, 2017.

SANTOS, Emelynne Gabrielly de Oliveira; BARBOSA, Isabelle Ribeiro; SEVERO, Ana Kalliny Sousa. Análise espaço-temporal da mortalidade por suicídio no Rio Grande do Norte, Brasil, no período de 2000 a 2015. **Ciência & Saúde Coletiva**, v. 25, p. 633-643, 2020.

SCHNITMAN, Gabriel et al. Taxa de mortalidade por suicídio e indicadores socioeconômicos nas capitais brasileiras. **Revista Baiana de Saúde Pública**, v. 34, n. 1, p. 46-46, 2010.

SCHWARZ, Gideon. Estimating the dimension of a model. **The annals of statistics**, p. 461-464, 1978.

SHAPIRO, Samuel Sanford; WILK, Martin B. An analysis of variance test for normality (complete samples). **Biometrika**, v. 52, n. 3/4, p. 591-611, 1965.

STASINOPOULOS, Mikis D. et al. **Flexible regression and smoothing: using GAMLSS in R**. CRC Press, 2017.

TAYLOR, Richard et al. Socio-economic differentials in mental disorders and suicide attempts in Australia. **The British Journal of Psychiatry**, v. 185, n. 6, p. 486-493, 2004.

VOUDOURIS, Vlasios et al. Modelling skewness and kurtosis with the BCPE density in GAMLSS. **Journal of Applied Statistics**, v. 39, n. 6, p. 1279-1293, 2012.