Universidade Federal de São Carlos
Centro de Ciências Exatas e de Tecnologia
Departamento de Estatística

Modelagem da Mortalidade por COVID-19 nos Países Europeus

Leonardo de Salles Amaral

Trabalho de Conclusão de Curso

Universidade Federal de São Carlos
Centro de Ciências Exatas e de Tecnologia
Departamento de Estatística

Modelagem da Mortalidade por COVID-19 nos Países Europeus

Leonardo de Salles Amaral
Orientador: Prof. Dr. Gustavo Henrique de Araujo Pereira

Trabalho de Conclusão de Curso apresentado como parte dos requisitos para obtenção do título de Bacharel em Estatística.

São Carlos
Fevereiro de 2025

Leonardo de Salles Amaral

Modelagem da Mortalidade por COVID-19 nos Países Europeus

Este exemplar corresponde à redação final do trabalho de conclusão de curso devidamente corrigido e defendido por Leonardo de Salles Amaral e aprovado pela banca examinadora.

Aprovado em 14 de Fevereiro de 2025

Banca Examinadora:

- Prof. Dr. Gustavo Henrique de Araujo Pereira
- Prof. Dr. Michel Helcias Montoril
- Prof. Dr. Rafael Izbicki

Agradecimentos

Agradeço à Deus e aos orixás por ter me fortalecido em toda a minha caminhada, aos meus pais por me proporcionarem a oportunidade em conseguir estudar em uma universidade pública de altíssima qualidade, independente das circunstâncias.

Aos amigos que construí e que acreditaram mais em do que eu mesmo, em especial a república Dauhma que me acolheu quando ingressei na UFSCar em 2019 e se tornaram grandes referências, aos meu professores que me ensinaram e proporcionaram tanto crescimento profissional, ao meu orientador por me dado a honra e o privilégio em ser seu orientando, dando todo suporte para conduzir esse trabalho e por ter me mostrado o quanto pode ser e é legal estudar estatística. E me agradeço profundamente por acreditar que a obsessão pode vencer o talento, caso contrário, esse momento não chegaria jamais.

"Eu cheguei de muito longe
E a viagem foi tão longa
E na minha caminhada
Obstáculos na estrada
Mas enfim aqui estou."
(Erasmo Carlos)

Resumo

O presente estudo da modelagem da mortalidade por COVID-19 nos países europeus traz indicativos de como cada nação europeia se comportou ao longo da pandemia. O objetivo deste estudo é modelar a mortalidade utilizando o ajuste de modelos GAMLSS paramétricos, pois esses modelos permitem que a variável resposta não pertença exclusivamente à família exponencial, possibilitando uma ampla variedade de modelagens com diversas distribuições.

Devido à escala dos danos provocados pela COVID-19 em nível global, este estudo busca identificar variáveis que possam explicar as diferenças na taxa de mortalidade entre os países europeus. Considerando a sensibilidade do tema, é crucial atentar para variáveis que, mesmo com uma relação aparentemente modesta, possam exercer impacto significativo sobre a variável resposta, visto que variações de apenas 0,1% nas covariáveis podem refletir em milhares de vidas afetadas. Os resultados deste estudo podem ajudar organizações de saúde e governos a tomar medidas necessárias para combater e gerir futuras pandemias.

Palavras-chave: *pandemia da COVID-19, modelos GAMLSS paramétrico.*

Abstract

The present study on modeling COVID-19 mortality in European countries provides insights into how each European nation behaved throughout the pandemic. The objective of this study is to model mortality using parametric GAMLSS models, as these models allow the response variable to belong to distributions beyond the exponential family, enabling a wide range of modeling options with various distributions.

Due to the scale of the damage caused by COVID-19 globally, this study aims to identify variables that can explain the differences in mortality rates among European countries. Considering the sensitivity of the subject, it is crucial to pay attention to variables that, even with an apparently modest relationship, can have a significant impact on the response variable, as variations of only 0.1% in the covariates may reflect in thousands of lives affected. The results of this study can help health organizations and governments take the necessary measures to combat and manage future pandemics.

Keywords: *COVID-19 pandemic, parametric GAMLSS models.*

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

A COVID-19, doença provocada por uma variante do coronavírus, surgiu no final de 2019 e se espalhou rapidamente pelo mundo. Em 2020, a Organização Mundial da Saúde (OMS) declarou o surto como uma pandemia, que é uma epidemia que ganha escala global. A pandemia durou mais de três anos, tendo seu fim declarado em maio de 2023. Nesse período, foram registrados mais de 700 mil mortes pela doença no Brasil (Alves, 2023). A pandemia da COVID-19 sobrecarregou sistemas de saúde até mesmo em países com estruturas médicas mais robustas e trouxe ainda mais desafios para regiões onde o acesso à saúde já era difícil.

Tal feito tornou-se uma preocupação mundial para as famílias, o governo e as empresas devido aos efeitos adversos na população e na força de trabalho global. Nos deparamos com inúmeros estudos que relataram taxas de mortalidade resultantes do impacto da pandemia usando diversas variáveis como os fatores socioeconômicos e culturais, dentre os estudos, destacam-se Sorci *et al.* (2020) e Canatay *et al.* (2021).

De maneira sucinta, a COVID-19 é uma doença infecciosa causada pelo coronavírus SARS-CoV-2 e tem como principais sintomas febre, cansaço e tosse seca (OMS, 2021). Outros sintomas menos comuns e que podem afetar alguns pacientes são: perda de paladar ou olfato, congestão nasal, conjuntivite, dor de garganta, dor de cabeça, dores nos músculos ou juntas, diferentes tipos de erupção cutânea, náusea ou vômito, diarreia, calafrios ou tonturas.

O tratamento para a COVID-19 varia de acordo com a gravidade da doença e o risco de agravamento da condição, incluindo a idade da pessoa e problemas de saúde característicos de cada grupo de risco. Portanto, os tratamentos devem ser decididos individualmente entre o paciente e o profissional de saúde que cuida dele.

Durante esse período, as prioridades se concentraram em medidas de prevenção e controle de infecções, treinamento de profissionais de saúde, atividades para alcançar comunidades afastadas, apoio à saúde mental e tratamento hospitalar para pacientes em estado grave.

No Brasil, a resposta à pandemia da COVID-19 foi a maior operação em 30 anos de história do Sistema Único de Saúde (SUS) no país (MSF, 2021). A assistência médica aos pacientes de COVID-19 foi oferecida em todos os níveis, assim como foi fornecido apoio de saúde mental para pacientes e profissionais de saúde. Também foram ministrados treinamentos para aprimoramento de protocolos e fluxos de pacientes. Adicionalmente, foi dada ênfase especial ao engajamento comunitário, com atividades de promoção de saúde e diagnósticos, utilizando em grande medida testes rápidos de antígeno.

A pesquisa de medicamentos para o tratamento da doença e o desenvolvimento de novas ferramentas, incluindo vacinas, possibilitaram maior controle sobre a COVID-19 globalmente. No entanto, a doença continua sendo objeto de estudo, principalmente em relação aos efeitos prolongados da COVID-19 (Pfizer, 2022).

Desenvolvidas a um ritmo sem precedentes, as vacinas surgiram no fim de 2020 e logo mudaram a trajetória do combate à pandemia, pelo menos para os países ricos que começaram a administrá-las em larga escala. Países de baixa e média renda sofreram com vacinações reduzidas, já que os países ricos monopolizavam os estoques de imunizantes (Butantan, 2021). A Campanha de Acesso de MSF (Médicos Sem Fronteiras) foi fortemente ativa nesse ponto, enfatizando a necessidade de distribuição equitativa em todo o mundo e pressionando fortemente por mecanismos para ampliação do acesso aos imunizantes e outras ferramentas importantes para o combate à COVID-19 (MSF, 2021).

Embora a epidemia de SARS-CoV-2 tenha se espalhado por todo o mundo, houve muita preocupação com a taxa de mortalidade que a infecção induziu entre a população. Para compreender seu impacto mundial, foi necessário um estudo demográfico sobre os países em estudos, compilando dados sobre demografia, economia e regimes políticos. Com tal abordagem, descobrimos que as trajetórias temporais da taxa de letalidade variam muito entre os países, o que era esperado, visto que os países apresentam políticas internas e culturas diferentes (de Moraes Bernal *et al.*, 2020).

Portanto, ao longo deste período algumas políticas públicas foram discutidas, elaboradas e colocadas em prática. Após debates com especialista, um dos pontos importantes foram estabelecer o ***lockdown***, que consiste em uma restrição severa ao deslocamento

de pessoas, fazendo com que reduzisse o risco de exposição ao vírus, não aumentando de forma exponencial os casos. Além disso, continuou-se usando medidas como o uso de máscaras, mantendo a higiene das mãos com álcool em gel, deixando os ambientes bem ventilados sempre que possível, evitando aglomerações e reduzindo ao máximo o contato próximo entre muitas pessoas, principalmente em espaços fechados (Ornelas, 2020). Outro ponto de destaque para que as pessoas voltassem a realizar seus compromissos diários foram as pesquisas para o desenvolvimento da vacina e políticas públicas para que todos os países tivessem recursos para vacinar toda sua população. Agora sabemos que a vacinação foi bastante importante para que a vida voltasse ao que era antes deste período histórico (Passarelli-Araujo *et al.*, 2022).

Neste trabalho, propomos uma abordagem via Modelos de Regressão Paramétricos para estudar as covariáveis que impactam a taxa de mortes por paises europeus da COVID-19. Variáveis demográficas, econômicas e políticas serão avaliadas. Além disso, serão usadas técnicas de seleção de variáveis e de análise de diagnostico para a obtenção de um modelo que ajuste adequadamente a taxa de morte por país.

Utilizando-se o modelo ajustado na etapa anterior do trabalho, obteremos o resíduo quantílico (Dunn e Smyth, 1996) para cada país. A partir desses resíduos, conseguiremos identificar países que apresentaram taxa de mortes inferior e superior ao esperado considerando suas características demográficas, econômicas e políticas.

Decidimos trabalhar apenas com um continente porque em uma análise prévia percebemos que o efeito das covariáveis variava bastante de acordo com o continente. E optamos pela Europa por dois motivos. O primeiro é que, por ser um continente sem países extremamente pobres, o grau de subnotificação de mortes por Covid é menor que em outros continentes. Além disso, pelo mesmo motivo, é mais razoável imaginarmos que um único modelo descreva a taxa de mortes por Covid em todo continente.

Este trabalho está organizado da seguinte maneira. No Capítulo 2, apresentamos a estrutura e os componentes de um Modelo GAMLSS paramétrico, procedimentos para estimação dos parâmetros pelo método da máxima verossimilhança, testes de hipóteses, intervalo de confiança e também outras metodologias capazes de selecionar o problema em estudo. Em seguida, o Capítulo 3 consiste no tratamento e análise descritiva dos dados. No Capítulo 4, discutimos sobre a criação do modelo proposto e como os países europeus se comportaram em relação a taxa de mortalidade, considerando questões demográficas, sociais e políticas. Por fim, no Capítulo 5, apresentamos uma síntese das conclusões

obtidas em relação ao objetivo em estudo, com base na metodologia proposta.

Capítulo 2

Modelos GAMLSS paramétricos

Neste capítulo serão discutidos os modelos de regressão paramétricos para o problema em estudo. Abordaremos uma classe de modelos de regressão paramétricos denominada GAMLSS paramétrico (Rigby e Stasinopoulos, 2005). Além disso, falaremos de um caso particular do modelo e abordaremos aspectos inferenciais e de diagnóstico do GAMLSS paramétrico.

2.1 Especificação do modelo

Os modelos GAMLSS são uma extensão dos modelos lineares generalizados (MLGs). Essa extensão, nos trouxe um leque de diferentes oportunidade para modelar os mais variados tipos de dados. Uma característica interessante da classe de modelos GAMLSS é que ela nos permite modelar todos os parâmetros da distribuição da variável Y em função das variáveis preditoras e não apenas sua média, como é feito nos MLGs. Outro ponto em destaque é que não é preciso que a variável resposta pertença exclusivamente à família exponencial, o que abre inúmeras possibilidades de modelagens com as mais diversas distribuições. Além disso, os modelos GAMLSS nos permite a flexibilidade em adicionar termos não parámetricos nos modelos. No entanto, a adição de termos não paramétricos reduz a interpretabilidade do modelo GAMLSS. Assim, usaremos nesse trabalho uma subclasse desses modelos denominada GAMLSS paramétricos. Considere que a variável resposta Y_i tem uma distribuição de probabilidade qualquer com até 4 parâmetros denotados, por θ_{i1}, θ_{i2}, θ_{i3} e θ_{i4}.

A partir disso o modelo GAMLSS paramétrico é definido como

$$\begin{cases} \eta_{i1} &= g_1(\theta_{i1}) = x_{i1}^\top \beta_1, \\ \eta_{i2} &= g_2(\theta_{i2}) = x_{i2}^\top \beta_2, \\ \eta_{i3} &= g_3(\theta_{i3}) = x_{i3}^\top \beta_3, \\ \eta_{i4} &= g_4(\theta_{i4}) = x_{i4}^\top \beta_4, \end{cases} \tag{2.1}$$

em que $\beta_1 = (\beta_{11}, \beta_{21}, \ldots, \beta_{1p_1})^\top, \beta_2(\beta_{12}, \beta_{22}, \ldots, \beta_{2p_2})^\top, \beta_3 = (\beta_{13}, \beta_{23}, \ldots, \beta_{3p_3})^\top$ e $\beta_4 = (\beta_{14}, \beta_{24}, \ldots, \beta_{4p_4})^\top$ são os vetores de parâmetros desconhecidos, $x_{i1} = (x_{i11}, x_{i21}, \ldots, x_{i1p_1})^\top, x_{i2} = (x_{i12}, x_{i22}, \ldots, x_{i2p_2})^\top, x_{i3} = (x_{i13}, x_{i23}, \ldots, x_{i3p_3})^\top$ e $x_{i4} = (x_{i14}, x_{i24}, \ldots, x_{i4p_4})^\top$ são constantes que representam os valores das variáveis preditoras para cada um dos parâmetros da distribuição e por fim, g_1, g_2, g_3 e g_4 são funções de ligação estritamente monótonas e duplamente diferenciáveis. Além disso, temos que η_{i1}, η_{i2}, η_{i3} e η_{i4} são preditores lineares utilizados para modelar diferentes aspectos da distribuição da variável resposta.

Em muitas aplicações práticas, é comum utilizar no máximo quatro parâmetros que geralmente representam a posição, a escala, a assimetria e a curtose. Nos modelos, os dois primeiros parâmetros θ_{i1} e θ_{i2}, são frequentemente referidos na literatura como parâmetros de posição (ou locação) e escala, respectivamente. Já os dois últimos parâmetros θ_{i3} e θ_{i4} são conhecidos como parâmetros de forma.

2.2 Estimação dos parâmetros

A estimação dos parâmetros nos GAMLSS paramétricos é realizado pelo método da máxima verossimilhança. Entretanto, assim como para MLGs, não há soluções analíticas para a maximização da função de verossimilhança e, portanto, a solução deve ser obtida a partir de algoritmos numéricos.

Rigby e Stasinopoulos (2005) propuseram algoritmos numéricos para a estimação dos parâmetros devido as suas complexidades analíticas. Dentre esses métodos, temos o algoritmo CG, que é baseado na generalização do algoritmo de Cole e Green (1992). Além disso, temos o algoritmo RS, que é basicamente uma generalização do algoritmo que foi apresentado e proposto por Rigby e Stasinopoulos (1996). Mais detalhes desse algoritmo podem ser vistos em Rigby e Stasinopoulos (2005).

Os modelos GAMLSS estão implementados na linguagem R (R Core Team, 2024) no pacote GAMLSS (Stasinopoulos e Rigby, 2008). Esse pacote e os pacotes associados desenvolvidos pelos mesmos autores contém as principais funções de ajuste dos modelos GAMLSS. Para utilizarmos esse pacote de maneira correta, devemos seguir algumas etapas, como especificar a distribuição da variável resposta, as funções de ligação e as covariáveis utilizadas em cada um dos submodelos. O pacote permite também a realização de análise de diagnóstico usando o resíduo que abordaremos na Seção 2.5.

2.3 Teste de hipóteses

Dentre a gama de testes de hipóteses, abordaremos a realização de maneira breve de dois dos principais testes de hipóteses para nosso problema. Assim, destacam-se o teste de Wald e teste da Razão de Verossimilhanças para os parâmetros dos modelos GAMLSS parámetricos. Caso nosso interesse seja verificar se um dos parâmetros é ou não diferente de zero, neste caso, o teste mais simples é o teste de Wald. Caso contrário, se estamos interessados em verificar se vários parâmetros são simultaneamente iguais a zero, o teste mais simples é o teste da Razão de verossimilhanças. A seguir, apresentaremos detalhes destes testes, como definição de hipóteses e estatística teste.

2.3.1 Teste de Wald

Em modelos paramétricos como o GAMLSS paramétrico, frequentemente desejamos testar se um dos parâmetros β_{jk} é diferente de zero ou não, ou seja, desejamos testar se a hipótese H_0: $\beta_{jk} = 0$ vs H_1: $\beta_{jk} \neq 0$. Em outras palavras, estamos interessados em investigar se o k-ésimo parâmetro da variável resposta é o ou não função da j-ésima variável preditora.

A estatística de teste, denominada estatística de Wald, é dada por

$$Q_W = \frac{\hat{\beta}_{jk}}{\sqrt{\widehat{\text{Var}(\beta_{jk})}}}$$

em que $\widehat{\beta_{jk}}$ é o estimador de máxima verossimilhaça de β_{jk} e $\widehat{\text{Var}(\beta_{jk})}$ é o estimador da variância assintótica do estimador de máxima verossimilhança de β_{jk}, calculado a partir da inversa da matriz de informação de Fisher. Se a hipótese nula for verdadeira, Q_W tem distribuição assintótica normal padrão.

2.3.2 Teste da Razão de Verossimilhança

Outra ferramenta importante na comparação de diferentes modelos GAMLSS paramétrico é o teste da razão de verossimilhança. Este teste visa avaliar a adequabilidade de dois modelos ao comparar a verossimilhança maximizada de um modelo completo, sob a hipótese nula, com a verossimilhança maximizada de um modelo sem restrições. Em modelos GAMLSS paramétricos, este teste é utilizado para verificar se a adição de parâmetros adicionais melhora significativamente o ajuste do modelo aos dados.

Para testar hipóteses a respeito dos parâmetros do modelo GAMLSS paramétrico, é necessário definir algumas partições de vetores e matrizes. Suponha que temos interesse em testar hipóteses relacionadas a um subconjunto dos vetores de parâmetros β_1, β_2, β_3 e β_4. Assim, esses vetores podem ser escritos como $\beta_1 = (\beta_{A1}^T, \beta_{B1}^T)^T$, $\beta_2 = (\beta_{A2}^T, \beta_{B2}^T)^T$, $\beta_3 = (\beta_{A3}^T, \beta_{B3}^T)^T$ e $\beta_4 = (\beta_{A4}^T, \beta_{B4}^T)^T$ em que β_{A1}, β_{A2}, β_{A3} e β_{A4} são vetores que contêm os parâmetros de interesse de dimensões, respectivamente, q_1, q_2, q_3 e q_4 e $0 \leq q_1 \leq p_1$, $0 \leq q_2 \leq p_2$, $0 \leq q_3 \leq p_3$ e $0 \leq q_4 \leq p_4$. Dessa forma, as hipóteses do teste são:

$$\begin{cases} H_0 : \beta_{A1} = \beta_{A1}^{(0)}, \beta_{A1} = \beta_{A2}^{(0)}, \beta_{A3} =_{A3}^{(0)}, \beta_{A3} = \beta_{A4}^{(0)}; \\ H_1 : \text{violação de pelo menos uma das igualdades de } H_0 \end{cases} \tag{2.2}$$

onde $\beta_{A1}^{(0)}$, $\beta_{A2}^{(0)}$, $\beta_{A3}^{(0)}$ e $\beta_{A4}^{(0)}$ são vetores contendo valores específicos para os parâmetros de interesse. A estatística da razão de verossimilhanças é dada por

$$Q_{RV} = 2\left[l(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4) - l(\tilde{\beta}_1, \tilde{\beta}_2, \tilde{\beta}_3, \tilde{\beta}_4)\right]$$

em que $l(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4)$ é o logaritmo da função de verossimilhança aplicado no estimador de máxima verossimilhança de β_{jk} do modelo completo e $l(\tilde{\beta}_1, \tilde{\beta}_2, \tilde{\beta}_3, \tilde{\beta}_4)$ é o logaritmo da função de verossimilhança aplicado no estimador de máxima verossimilhança do modelo com restrições nos vetores de parâmetros (sob H_0).

2.4 Intervalos de Confiança

Assim como os testes de hipótese, outra opção que pode ser explorada são os intervalos de confiança. Para obtermos intervalos de confiança para os parâmetros de um modelo GAMLSS, podemos usar métodos como o método da razão de verossimilhança, o método

de Wald ou métodos baseados em bootstrap. De maneira sucinta, usando o método de Wald, como estamos interessados em apresentar o intervalo de confiança para determinado parâmetro em estudo, temos que para um nível de confiança $1-\alpha$, o intervalo de confiança para cada parâmetro β_{jk} é dado por

$$\hat{\beta}_{jk} \pm Z_{\alpha/2}\sqrt{\widehat{\text{Var}(\beta_{jk})}},$$

em que $Z_{\alpha/2}$ é o quantil (1 - $\alpha/2$) da distribuição normal padrão.

2.5 Análise de Diagnóstico

Os resíduos fornecem informações valiosas para diagnosticar e avaliar a qualidade do ajuste de um modelo, ajudando a verificar se as suposições do modelo foram atendidas. Diversas ferramentas gráficas são empregadas para identificar discrepâncias entre o modelo ajustado e as observações coletadas.

O resíduo mais utilizado na análise de diagnóstico de modelos GAMLSS paramétricos é o resíduo quantílico aleatorizado (Dunn e Smyth, 1996). Esse resíduo é interessante por ter distribuição assintoticamente normal padrão se o modelo é o correto e se são utilizados estimadores consistentes para os parâmetros. Quando a variável resposta é contínua, o cálculo do resíduo não envolve aleatorização e o resíduo pode ser chamado apenas de resíduo quantílico.

Para o modelo GAMLSS paramétrico, o resíduo quantílico é dado por

$$r_i^q = \Phi^{-1}\left\{F(y_j;\hat{\theta}_{j1},\hat{\theta}_{j2},\hat{\theta}_{j3},\hat{\theta}_{j4})\right\}, \tag{2.3}$$

em que $\Phi^{-1}(\cdot)$ e $F(\cdot)$ representam as funções de distribuição acumulada da distribuição normal padrão e da distribuição da variável resposta.

2.6 Modelos Lineares Generalizados Duplos (DGLM)

Introduzidos por Smyth (1989), os Modelos Lineares Generalizados Duplos (DGLM) são utilizados quando os dados apresentam dispersão não homogênea ou quando há interesse em modelar tanto a média quanto um parâmetro de dispersão da resposta, ampliando a capacidade de ajuste e explicação de fenômenos mais complexos. Os DGLM são um

caso particular dos Modelos GAMLSS paramétricos, ambos buscam compreender como os diferentes parâmetros de uma distribuição são influenciados por variáveis explicativas. No entanto, nos DGLM, a distribuição da variável resposta pertence à família exponencial.

Uma variável aleatória Y pertence à família exponencial se pudermos reescrever sua função de densidade de probabilidade da seguinte forma:

$$f(y;\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{\phi} + C(y,\phi)\right\} I_A(y), \tag{2.4}$$

em que o valor esperado de Y é dado por $E(Y) = \mu = b'(\theta)$, a função de variância associada à média é $V(\mu) = b''(\theta(\mu))$, e a variância de Y pode ser obtida por meio de $\text{Var}(Y) = \phi V(\mu)$. Além disso, A é o suporte de Y, ou seja, um conjunto que não depende de θ ou ϕ. As principais distribuições de probabilidade que pertence à família exponencia são a normal, gama, gaussiana inversa, Poisson e binomial.

Os Modelos Lineares Generalizados Duplos (DGLM) consistem em dois submodelos. O primeiro submodelo é responsável pela modelagem da média, enquanto o segundo se encarrega de ajustar o parâmetro de dispersão de uma variável de resposta. Eles podem ser definidos da seguinte maneira:

- O primeiro submodelo relaciona a **média** ($E(Y)$) com as variáveis explicativas.
- O segundo submodelo ajusta o parâmetro de **dispersão** (ϕ), modelando a variabilidade da variável de resposta.

Além disso, supomos que, seja $y_1, y_2, \ldots, y_n$ um conjunto de variáveis aleatórias independentes que seguem uma distribuição da família exponencial. Os DGLM pertencem a família exponencial e pelos seguintes componentes sistemáticos:

$$\begin{cases} g(\mu_i) = \eta_i, \\ h(\phi_i) = \zeta_i, \end{cases} \tag{2.5}$$

onde $\eta_i = x_{i1}^T\beta$ e $\zeta_i = x_{i2}^T\gamma$ são os preditores lineares, $\beta = (\beta_1, \beta_2, \ldots, \beta_{p1})^T$ e $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_{p2})^T$ são os vetores de parâmetros desconhecidos, e $y = (y_1, y_2, \ldots, y_n)^T$. Os vetores $x_{i1} = (x_{i11}, x_{i21}, \ldots, x_{ip_11})^T$ e $x_{i2} = (x_{i12}, x_{i22}, \ldots, x_{ip_22})^T$ representam os valores das variáveis explicativas para os dois submodelos, com $p_1 + p_2 < n$, e as funções $g(\cdot)$ e $h(\cdot)$ são estritamente monótonas e duas vezes diferenciáveis, representando as funções de ligação para os submodelos de μ e ϕ, respectivamente.

Os DGLM têm sido usados na prática em diversas áreas, como pode ser visto em Haatveit *et al.* (2023) e Liu *et al.* (2023). Além disso, trabalhos teóricos relacionados ao modelo tem sido desenvolvidos como Paula (2013) e Cavalaro e Pereira (2022).

2.7 Outras metodologias

Com o avanço dos estudos estatísticos e a crescente complexidade dos problemas práticos, além dos modelos GAMLSS paramétricos, surgiram diferentes mecanismos e abordagens para atender às diversas demandas em análise de dados. Essas soluções incluem modelos não-paramétricos, métodos semiparamétricos e técnicas baseadas em aprendizado de máquina, cada uma oferecendo ferramentas específicas para lidar com estruturas de dados complexas, heterogeneidade e dependências que não podem ser adequadamente capturadas pelos modelos tradicionais. Esse desenvolvimento reflete a constante evolução da estatística em resposta aos desafios do mundo real.

Dentre essas metodologias, uma técnica amplamente utilizada é a Máquinas de Vetores de Suporte (SVMs), que foi introduzida por (Vapnik, 1995). Elas visam criar classificadores com boa capacidade de generalização, ou seja, capazes de prever novos casos corretamente com base no aprendizado. Os SVMs determinam um hiperplano ótimo que minimiza o risco empírico e a Dimensão VC, maximizando a margem de separação entre as classes e minimizando o risco estrutural. Além disso, podem ser utilizados em diversos casos devido à sua eficiência na classificação e regressão. Elas são frequentemente aplicadas em situações como classificação de texto (Joachims, 1998) e reconhecimento de voz (Maitra *et al.*, 2015), com isso, nota-que essa metodologia é valiosa, conseguindo lidar com grandes conjuntos de dados e podendo funcionar bem em espaços de alta dimensionalidade, sendo eficaz em casos onde é importante encontrar um equilíbrio entre precisão e complexidade do modelo (Schölkopf e Smola, 2002).

As SVMs e outras técnicas de aprendizado de máquina poderiam ser usadas neste trabalho. Porém, conforme será visto no Capítulo 4 foi possível obter um bom ajuste da variável resposta usando um DGLM. Além disso, o DGLM gera modelos mais interpretáveis do que as principais técnicas de aprendizado de máquina. Como um dos principais motivos é entender a relação entre as covariáveis disponíveis, por esse motivo e pelo bom ajuste obtido, optamos neste trabalho modelar a taxa de mortes por milhão na Europa usando um DGLM.

Capítulo 3

Análise exploratória de dados

Este capítulo aborda às principais práticas relacionados à coleta e ao tratamento de dados, bem como à análise descritiva. Na seção sobre coleta e tratamento de dados, será enfatizado a importância de um planejamento para assegurar a qualidade e a relevância das informações obtidas. Em seguida, na análise descritiva, o foco será em técnicas para resumir e interpretar os dados, utilizando principalmente técnicas gráficas para a obtenção de conclusões preliminares relacionadas com o objetivo do trabalho.

3.1 Coleta e tratamento dos dados

Iniciado em 29 de janeiro de 2020, o conservatório de dados, começou a rastrear o coronavírus, oferecendo as estatísticas globais mais precisas e oportunas durante esse período desafiador (Tracker, 2024). Após o fim da pandemia da COVID-19, aproximadamente dois anos, em 13 de abril de 2024, o Coronavirus Tracker deixou de ser atualizado devido à dificuldade de fornecer totais globais estatisticamente válidos, uma vez que a maioria dos países cessou os relatórios. No entanto, os dados históricos continuaram acessíveis.

Ao longo da pandemia, o banco de dados serviu como um repositório em tempo real e auxiliou fortemente como fonte de estudo para muitos pesquisadores. Como nosso principal objetivo é propor uma modelagem capaz de analisar o comportamento dos países europeus em relação à COVID-19, realizamos um estudo complementar de variáveis para compor nosso banco de dados. O banco de dados inicial é composto por 231 países diferentes, que estavam presentes no repositório. Essa etapa de composição do banco de dados foi preciso analisar do ponto de vista analítico quais covariáveis são coerentes serem selecionadas de acordo com o nosso interesse. Fixamos o ano de 2019 como referência para

buscarmos informações das covariavéis que podem a vir serem significativas em nosso modelo.

As covariáveis mencionadas estão presentes em Sorci *et al.* (2020) e Canatay *et al.* (2021). Para obtermos os dados utilizamos repositórios de dados online disponíveis em (in Data, 2024) e (Bank, 2024) para selecionarmos as variavéis que serão apresentadas posteriormente.

Em nosso estudo analítico, após a inclusão das variavéis selecionadas no banco de dados, percebemos que a base estruturada apresentava valores faltantes (missings) em excesso nas covariáveis. Por conta disso, optamos em não aplicar nenhuma técnica de imputação de dados e realizamos uma uma pré-seleção de covariáveis manual, buscando excluir covariáveis que possuiam um elevado número de missings e países que tinham observações faltantes em covariáveis com baixa proporção de missing. Essa estruturação foi executada via Excel, utilizando técnicas eficazes como filtros aplicado em cada variável em diferentes cenários de comparação.

Logo, como estamos interessados em estruturarmos um banco de dados que apresenta o máximo de covariáveis com informações preenchidas em relação aos países em estudo optamos por deixar covariáveis como o número de leitos hospitalares, gastos com educação, número de médicos (por 1000 habitantes), despesa total com saúde como parcela do produto interno bruto (PIB) nacional e taxa de mortalidade por diabetes mellitus entre ambos os sexos de fora da base por apresentarem missing excessivos em relação a muitos países.

Além disso, temos que os dados em relação às principais comorbidade como as taxas de mortalidades por poluição do ar, tabagismo, doenças por câncer, doenças respiratórias crônicas e doenças renais crônicas não estavam disponíveis para download. Não apenas, outras covariáveis como o índice de rigor do governo ***(lockdown)***, despesas com pandemia em relação ao PIB Per Capita e taxas de individualismo e coletivismo entre os habitantes não foram encontradas durante a pesquisa nos repositórios online e também ficaram de fora da estruturação prévia. Com isso, o banco de dados apresenta como covariáveis independentes e disponíveis para estudos a taxa de mortalidade por doenças cardiovasculares, taxa de mortalidade infantil, taxa de fertilidade, taxa da população urbana (%), densidade populacional, índice de desenvolvimento humano (IDH), esperança de vida, PIB Per Capita e taxa de turistas.

Por fim, ao longo desse processo, percebemos a estrutura que o banco de dados estava

ficando, dentre os 231 países inicias em estudo, considerando todos os continentes, 90 países ficaram de fora devido o excesso de missings nas covariáveis e 102 por não pertencerem ao continente europeu. Portanto, a base final contempla 39 países e 9 covariáveis que apresentam informações completas em relação aos países europeus que permaneceram na base.

3.2 Análise descritiva dos dados

O banco de dados final contém 39 países e as variáveis são apresentadas na Tabela 3.1

Indicador	**Tipo**
País	Nominal
Mortes: É o número de óbitos por COVID-19 dividido pela população do país multiplicado por 1 milhão de habitantes.	Contínua
TMC: Taxa de mortalidade por doenças cardiovasculares	Contínua
TMI: Taxa mortalidade infantil	Contínua
TF: Taxa de fertilidade	Contínua
TU: Taxa da população urbana	Contínua
DEN: Densidade populacional	Contínua
IDH: Índice de desenvolvimento humano	Contínua
EV: Esperança de vida	Contínua
PPC: PIB Per Capita	Contínua
TUR: Taxa de turistas	Contínua

Tabela 3.1: Descrição da variável resposta e covariáveis em estudo.

Note que todas as covariáveis são contínuas. Com a apresentação das variáveis em estudo finalizada, a Tabela 3.2 apresenta a estrutura do banco de dados em estudo.

PAÍSES	IDH	EV	PPC	...	TF	TMC	MORTES
Albânia	0.8	79.2825	11715.307	...	1.395	8090.849561	1258
Armênia	0.789	75.4386	12399.231	...	1.601	4998.54641	2953
Áustria	0.92	81.9077	43488.668	...	1.464	3654.518284	2486
Bielorrússia	0.81	74.216	19130.395	...	1.389	8080.127961	755
Bélgica	0.936	81.8311	41155.527	...	1.609	2502.303077	2946
Bulgária	0.813	75.0624	18840.764	...	1.58	10720.98084	5661
Croácia	0.866	78.7376	23590.277	...	1.47	5402.022949	4604
Chipre	0.901	81.397	30284.795	...	1.33	1740.23222	1116
Dinamarca	0.946	81.4337	48238.656	...	1.698	2100.755673	1511
Estônia	0.893	78.6693	28505.604	...	1.661	5841.31561	2270
Finlândia	0.939	81.8706	39874.8	...	1.351	3789.872005	2153
França	0.905	82.7315	39084.656	...	1.826	2145.470675	2556
Geórgia	0.816	73.4696	12631.664	...	2.021	6780.410848	4317
Alemanha	0.951	81.5584	46761.945	...	1.541	4095.032742	2181
Hungria	0.854	76.4543	27272.312	...	1.531	6485.753301	5106
Islândia	0.958	82.4042	43136.605	...	1.744	1883.970346	663
Irlanda	0.942	82.2586	59733.91	...	1.718	1792.859835	1891
Itália	0.899	83.552	35407.242	...	1.259	3648.955021	3260
Letônia	0.873	75.5337	25858.602	...	1.609	8301.670688	3630
Luxemburgo	0.925	82.1434	54751.6	...	1.343	1868.174406	1918
Malta	0.905	83.2065	33220.285	...	1.15	2606.484917	1993
Moldávia	0.773	70.9351	7078.23	...	1.78	7084.569466	3044
Montenegro	0.841	77.0396	20448.549	...	1.814	6258.705396	4532
Holanda	0.941	82.0455	48142.67	...	1.572	1919.885379	1336
Noruega	0.961	82.9552	85115.32	...	1.533	1884.417722	1204
Polônia	0.88	77.9272	29142.023	...	1.436	4725.238391	3196
Portugal	0.864	81.7007	28056.057	...	1.421	3324.732218	2773
Romênia	0.834	76.5079	23267.516	...	1.711	7305.042916	3622
Rússia	0.839	73.9332	25244.12	...	1.504	6583.472859	2762
Sérvia	0.812	76.704	15293.246	...	1.513	8042.017297	2087
Eslováquia	0.863	77.6851	27429.225	...	1.566	4313.07957	3887

PAÍSES	IDH	EV	PPC	...	TF	TMC	MORTES
Eslovênia	0.918	81.6042	30229.299	...	1.611	3768.854881	3417
Espanha	0.904	83.5321	34957.465	...	1.23	2465.077369	2606
Suécia	0.947	83.0524	45434.402	...	1.709	2786.416467	2680
Suíça	0.96	83.7802	62448.598	...	1.478	2429.80754	1647
Ucrânia	0.774	74.5364	10139.761	...	1.218	9263.073483	2603
Reino Unido	0.933	81.725	39113.01	...	1.632	2249.65579	3389
Grécia	0.89	81.2489	23869.793	...	1.337	4464.737573	3671
Lituânia	0.886	76.2119	28986.053	...	1.61	8034.78498	3718

Tabela 3.2: Estrutura do banco de dados.

De maneira panorâmica, apresentaremos a seguir os países europeus presentes em nosso estudo pelo mapa mundi, sendo possível visualizar de como foi o período da pandemia da COVID-19. Devido ao tamanho da Rússia, serão apresentados dois gráficos, em que no segundo foi excluída a Rússia para melhor visualização.

Os diferentes tons de azul apresentados na Figura 3.2 permitem uma análise visual clara do número de mortes por COVID-19 por milhão (mpm) para cada país. Os países com tons mais escuros de azul registram maiores números de óbitos, enquanto aqueles com tons mais claros apresentam números significativamente menores. Essa variação de cores ilustra a discrepância entre as nações em termos de mortalidade.

Os países com maior número de mortes estão concentrados no leste europeu, uma região menos desenvolvida da Europa. Entre os países com os maiores índices de mortes estão Bulgária (5661 mpm), Hungria (5106 mpm) e Croácia (4604 mpm). Por outro lado, os países com os menores números são Islândia (663 mpm), Bielorrússia (755 mpm) e Chipre (1116 mpm). Essa discrepância reflete diferenças nas condições socioeconômicas, políticas públicas e sistemas de saúde.

Ao analisar o impacto do tamanho da população no número de mortes, não se observa um indicativo direto dessa relação. Os países de menor população apresentam uma grande variabilidade no número de mortes por milhão. Entre os países de pequena população, destacam-se Islândia (663 mpm) e Chipre (1116 mpm) com baixos índices, Malta (1993 mpm) e Estônia (2270 mpm) com valores médios, e Montenegro (4532 mpm) e Letônia (3630 mpm) com altos índices de mortes por milhão. Por outro lado, entre os países

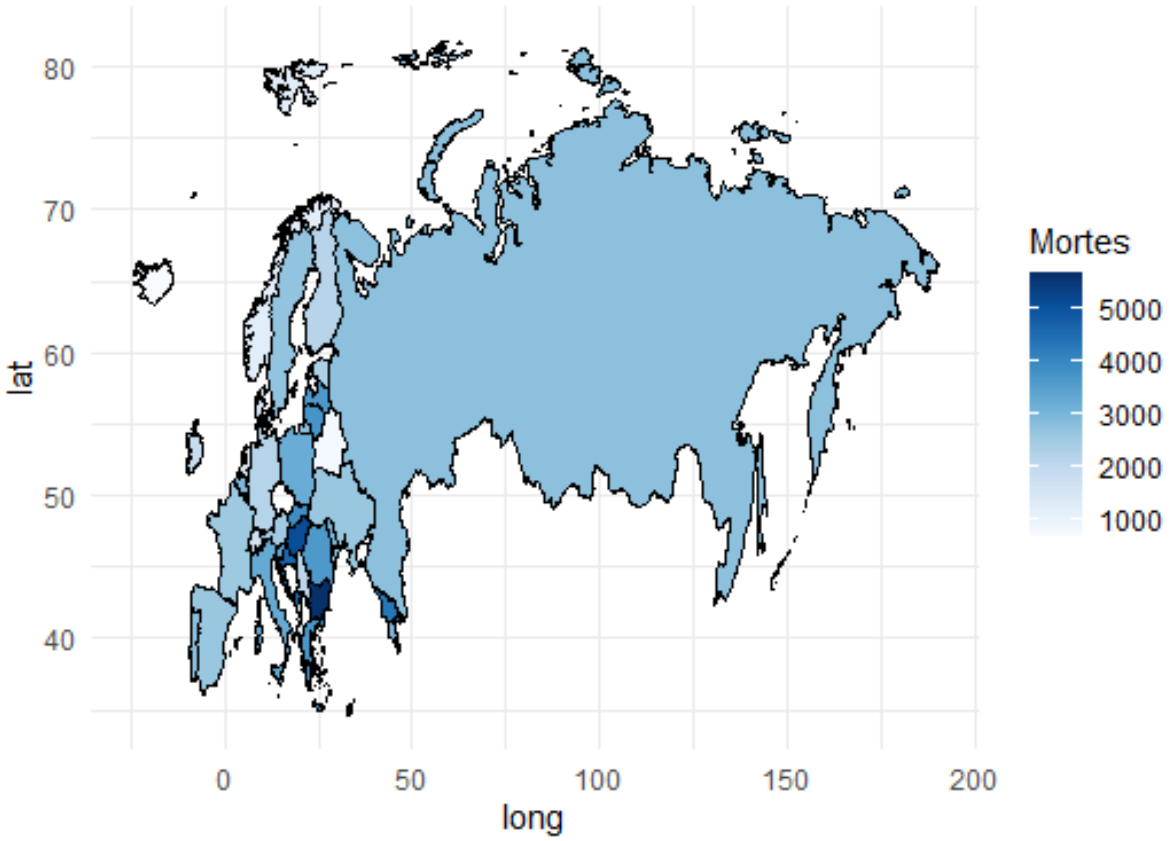

Figura 3.1: Mapa Mundi da mortalidade por milhão pela COVID-19 nos países europeus no presente estudo.

mais populosos da Europa, também há uma diversidade: Holanda (1336 mpm) apresenta baixa mortalidade, França (2556 mpm) registra mortalidade moderada, e Reino Unido (3389 mpm) está entre os países com alta mortalidade. No entanto, nenhum dos países mais populosos figura entre aqueles com os maiores números de mortes por milhão na Europa.

Com isso, temos que, embora o tamanho da população possa influenciar o número absoluto de mortes, outros fatores, como condições socioeconômicas, acesso a cuidados de saúde e políticas públicas, podem ser mais determinantes para explicar as variações na mortalidade entre os países. Portanto, uma investigação mais aprofundada é necessária para compreender completamente as causas por trás dessas diferenças.

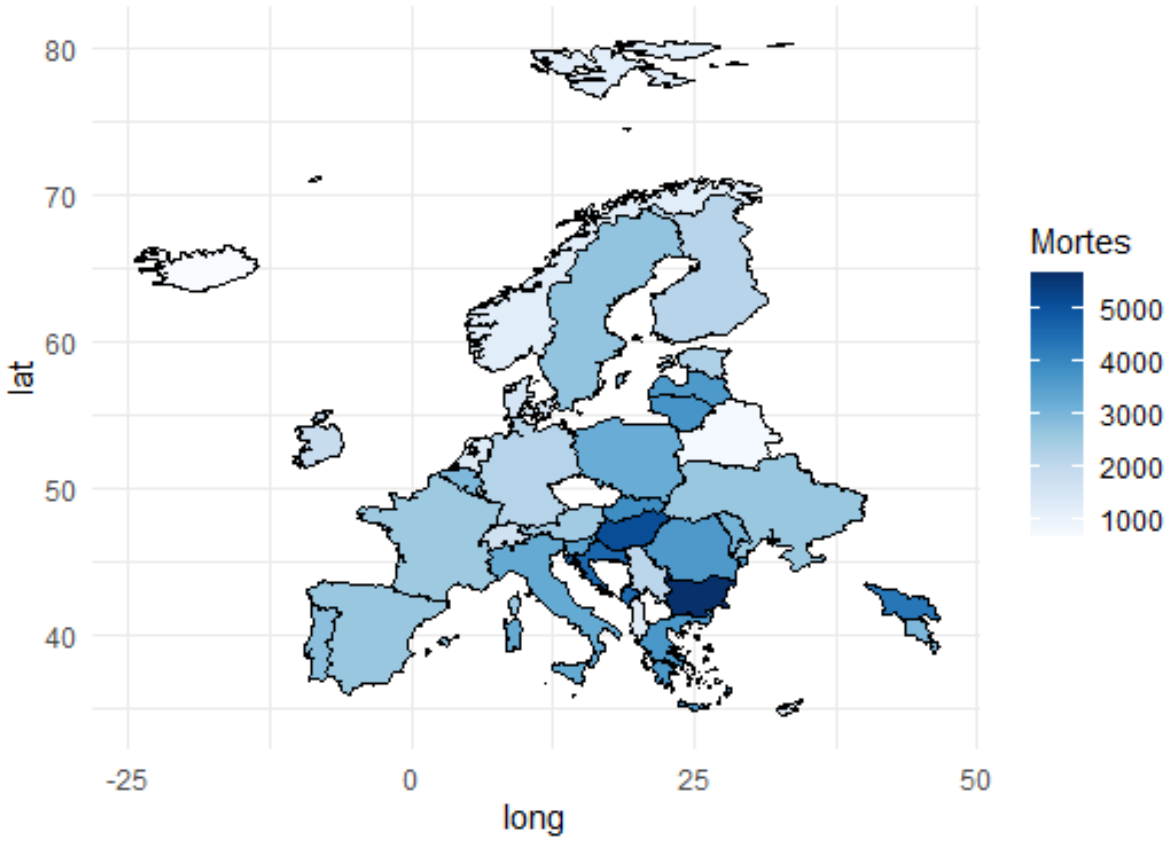

Figura 3.2: Mapa Mundi da mortalidade por milhão pela COVID-19 nos países europeus no presente estudo, exceto a Rússia.

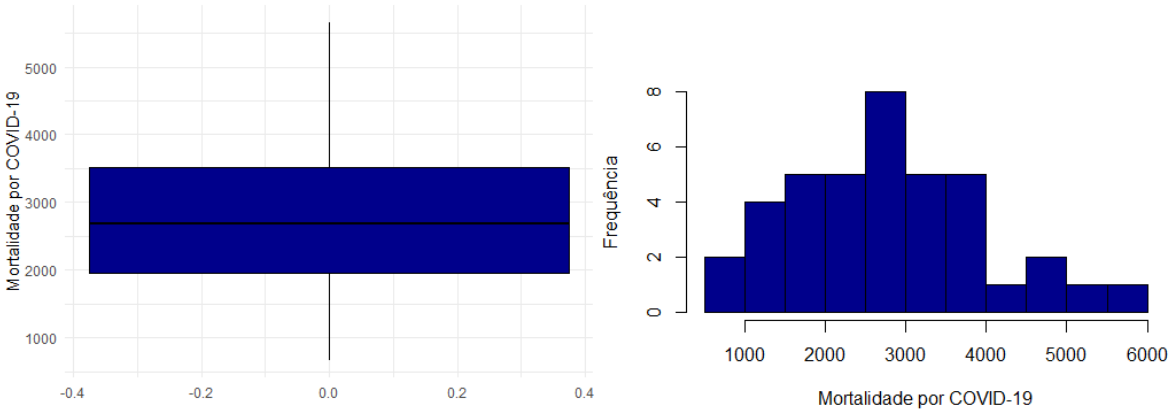

Figura 3.3: Box plot e histograma da variável resposta Mortalidade por 1 milhão de pessoas

A Figura 3.3 apresenta o boxplot e o histograma da taxa de mortes por COVID-19. Observa-se que a taxa parece apresentar leve assimetria à direita. A média da taxa é de 2680 mortes por COVID-19 para cada milhão de habitantes. Considerando que a assimetria amostral é pequena e que em um modelo de regressão consideramos a distribuição da

variável resposta dadas as covariáveis, é possível que um modelo GAMLSS paramétrico que assuma distribuição simétrica para a variável resposta ajuste de forma adequada o número de mortes por COVID-19 por milhão. Entretanto, nos gráficos apresentados analisamos a distribuição da variável resposta sem levar em conta as covariáveis. Por outro lado, nos modelos de regressão assumimos a distribuição dos dados considerando as covariáveis, avaliando os resíduos para verificar se uma distribuição simétrica é adequada para a variável resposta. Faremos essa avaliação na Seção 4.2.

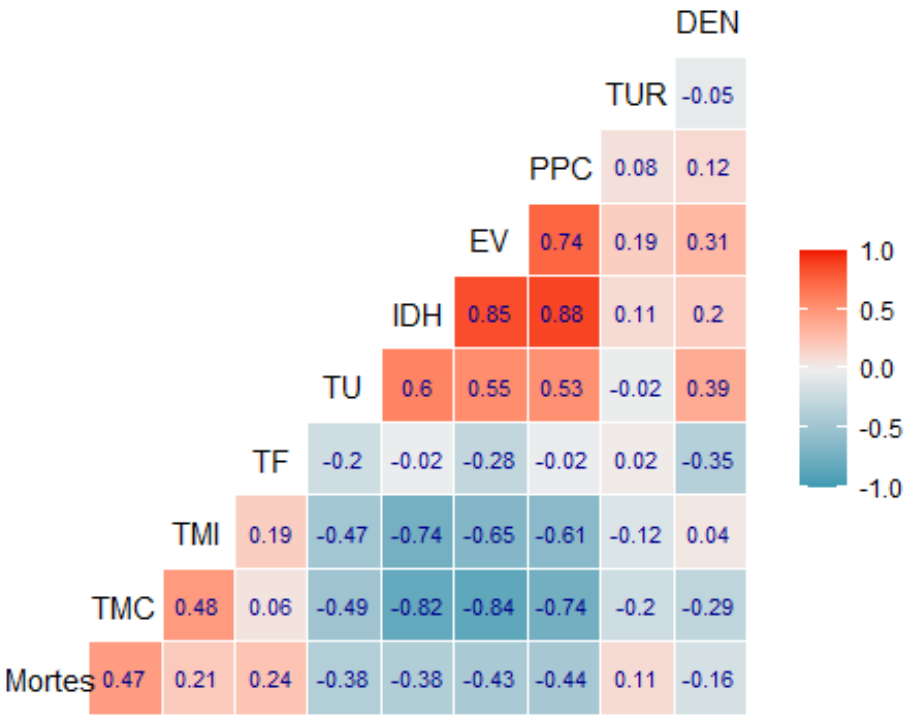

Figura 3.4: Matriz de correlação das variáveis em estudo

A Figura 3.4 apresenta as correlações entre as diferentes covariáveis e destas com a variável resposta usando as cores como indicativo das correlações. Observa-se que os tons das cores vermelhas indicam correlações positivas, enquanto os tons das cores azuis indicam correlações negativas.

Dentre as correlações presente na Figura 3.4, nota-se que há uma correlação de 0.47, sugerindo que a taxa de mortalidade por doenças cardiovasculares está moderadamente relacionada com a taxa de mortes por COVID-19. Com isso, temos um indicativo de que está covariável pode ter impactos importante na variável resposta.

Além disso, temos correlações positivas moderadas de 0.21 e 0.24, indicando que países com maiores taxa de mortalidade infantil e taxa de fertilidade tendem a ter mais mortes

por COVID-19. Tais fatores podem ser atribuídos ao fato de países mais pobres terem, em geral, maiores valores para essas variáveis. Por outro lado, a correlação positiva muito baixa de 0.11 está apontando que a taxa de turistas não está fortemente relacionado com o número de mortes por COVID-19.

Entretanto, com as correlações negativas moderadas de -0.38, -0.38, -0.43 e -0.44 temos um indicativo de que países com maiores taxas de população urbana, índice de desenvolvimento humano, esperança de vida ao nascer e Pib Per Capita, respectivamente, tendem a ter menos mortes por COVID-19. Isso pode estar relacionado aos países desenvolvidos que apresentam melhores condições de vida para sua população.

Há algumas covariáveis com correlação entre si, em módulo bem altas, sendo iguais ou superiores a 0,80, são elas: taxa de mortalidade por doenças cardiovasculares e índice de desenvolvimento humano, índice de desenvolvimento humano e esperança de vida ao nascer e esperança de vida ao nascer e pib per capita. Provavelmente nenhum par dessas variáveis entrarão ao mesmo tempo no modelo para evitar problemas de multicolinearidade (Montgomery *et al.*, 2006).

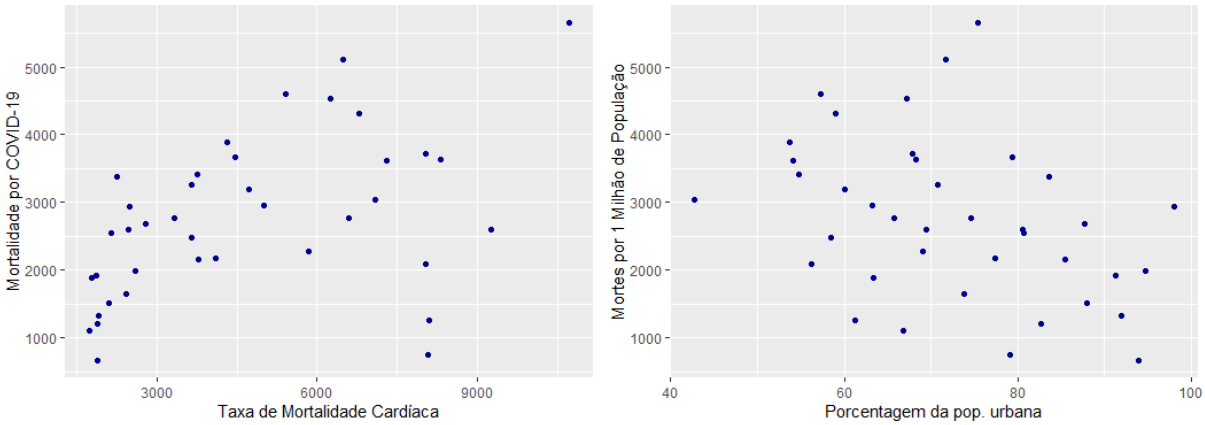

Figura 3.5: Diagrama de dispersão da mortalidade por COVID-19 em relação a taxa de mortalidade por doenças cardiovasculares (esquerda) e taxa da população urbana (direita).

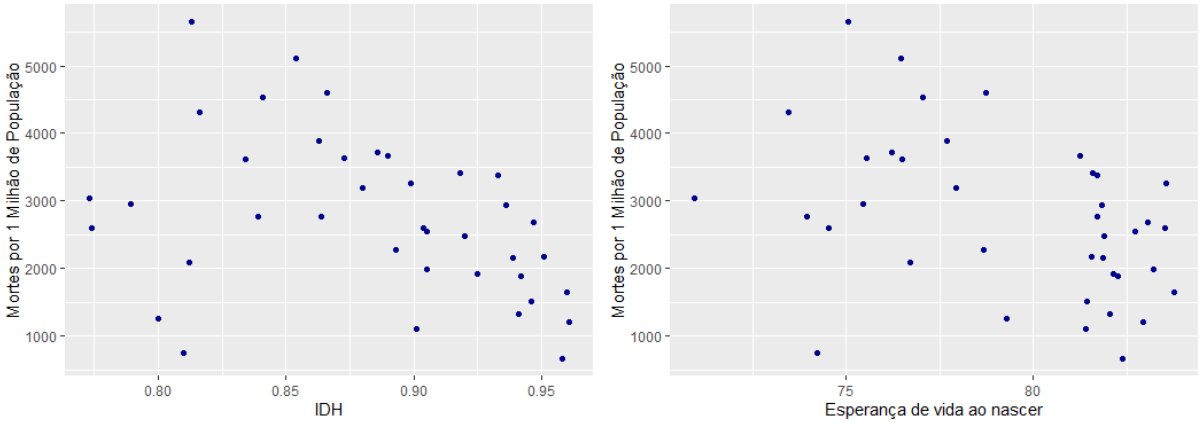

Figura 3.6: Diagrama de dispersão da mortalidade por COVID-19 em relação ao índice de desenvolvimento humano (esquerda) e esperança de vida ao nascer (direita).

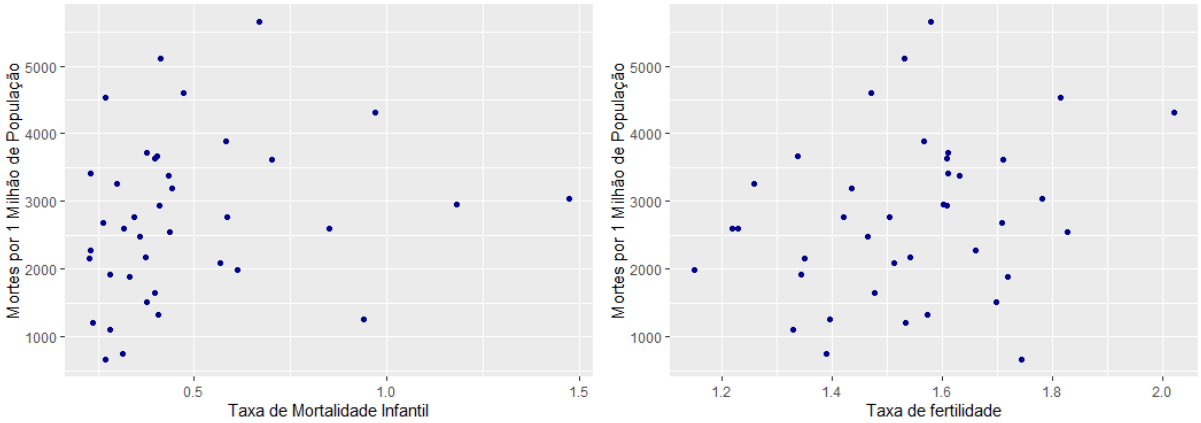

Figura 3.7: Diagrama de dispersão da mortalidade por COVID-19 em relação a taxa de mortalidade infantil (esquerda) e taxa de fertilidade (direita).

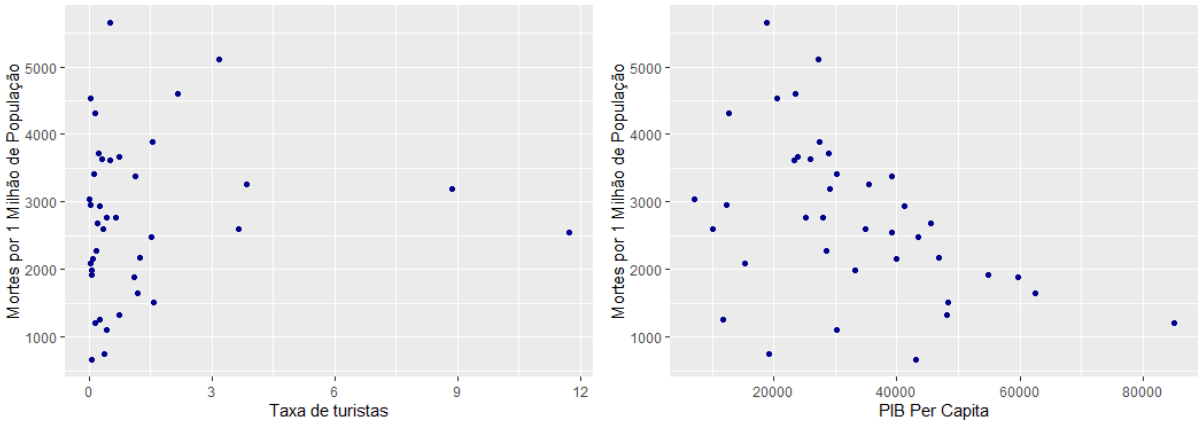

Figura 3.8: Diagrama de dispersão da mortalidade por COVID-19 em relação a taxa de turistas (esquerda) e PIB Per Capita (direita).

As Figuras 3.5 a 3.9 apresentam diagrama de dispersão envolvendo cada uma das

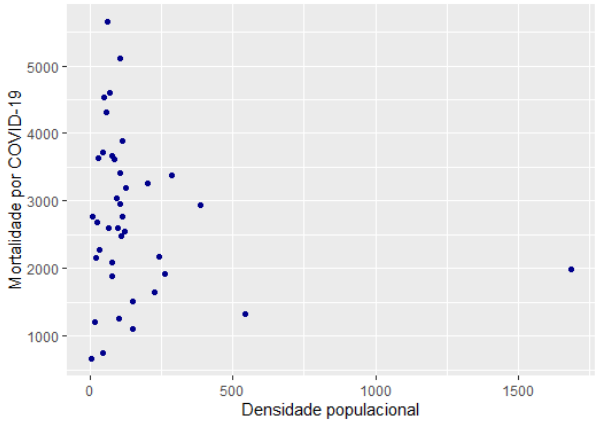

Figura 3.9: Diagrama de dispersão da mortalidade por COVID-19 em à densidade populacional.

covariáveis e a resposta. Além do já observado na Figura 3.3, conseguimos observar duas questões adicionais. O primeiro é que a variabilidade da variável resposta, mesmo fixando um valor de qualquer covariável é bastante alta. Além disso, a variância da variável resposta não parece ser constante para todos os valores de uma covariável. Isso fica bem evidente na Figura 3.5 gráfico da direita em que a variabilidade da resposta parece crescer com o valor da porcentagem da população urbana.

Considerando-se a matriz de correlações e os diagramas de dispersão, nota-se que, na Europa, as variáveis que apresentam valores mais altos para países desenvolvidos tendem a ter correlação negativa com a variável resposta. Por outro lado, as variáveis associadas a países menos desenvolvidos apresentam correlação positiva com a variável resposta. Isso sugere que a epidemia de COVID-19 foi mais grave nos países menos desenvolvidos da Europa. Esse fato pode estar relacionado a fatores como menor acesso a sistemas de saúde eficientes e condições socioeconômicas mais desfavoráveis. Essa hipótese será analisada com maior profundidade na continuidade deste trabalho.

Capítulo 4

Análise Inferencial

Para dar início à aplicação da modelagem, é importante destacar que já conduzimos a análise descritiva e exploratória de dados. Esses passos preliminares nos permitiram explorar e compreender as características principais dos dados, bem como identificar as possíveis relações entre as variáveis estudadas. Agora, estamos interessados em apresentar um modelo parcimonioso que seja capaz de resolver o problema proposto de maneira eficaz. Nossa meta é encontrar um modelo que, utilizando o menor número de covariáveis possível, consiga capturar a essência do fenômeno estudado.

Diversos modelos da classe GAMLSS paramétrico foram avaliados, levando em consideração covariáveis relevantes identificadas na análise descritiva. Dentre os modelos testados, o modelo linear generalizado duplo com distribuição normal com segunda parametrização (Rigby *et al.*, 2019) apresentou o melhor ajuste aos dados, conforme indicado pela análise de resíduos apresentada na Seção 4.2. Para a modelagem do parâmetro de média (μ), foi adotada a função de ligação identidade, garantindo uma relação linear direta entre a resposta e as covariáveis. Já para o parâmetro de dispersão(ϕ), optou-se pela função de ligação logarítmica, assegurando que os valores estimados para esse parâmetro fossem sempre positivos.

O processo de modelagem foi iniciado com a construção de modelos individuais para todas as variáveis, permitindo uma análise detalhada de cada uma isoladamente. Esse passo inicial foi crucial para identificar a influência de cada variável sobre o problema estudado. Em seguida, procedemos à criação de modelos em dupla, mantendo a variável que apresentou o menor valor-p nos modelos individuais. Esse método permitiu uma combinação otimizada das variáveis, assegurando que as mais significativas fossem preservadas e combinadas de forma a maximizar a precisão e a eficiência do modelo final.

Essa abordagem progressiva garantiu a seleção criteriosa das variáveis, visando sempre a construção de um modelo parcimonioso e robusto.

Optamos por avançar com o modelo duplo de menor valor-p, utilizando-o como base para iniciar a construção dos modelos triplos. Isso nos permitiu agregar valor incremental e significativo ao modelo, mantendo as variáveis mais relevantes identificadas na análise dupla. Após identificar o modelo triplo com o menor valor-p, prosseguimos para a criação dos modelos com quatro variáveis, focando em otimizar a relevância e a precisão dos nossos resultados. Esse processo iterativo garantiu que cada etapa da modelagem fosse fundamentada nas combinações de variáveis mais estatisticamente significativas, culminando na formulação de um modelo final mais eficiente.

Nessa etapa, identificamos os modelos mais coerentes com o nosso problema, avaliando a consistência dos sinais das estimativas com as correlações apresentadas anteriormente. Essa verificação assegurou a consistência e validade dos modelos, reforçando a relevância das variáveis selecionadas e a robustez das conclusões. Assim, garantimos que o modelo final fosse estatisticamente significativo e coerente com a lógica do problema estudado. Essa metodologia de modelos encaixados, assegurou que os modelos desenvolvidos fossem estatisticamente sólidos e adequados às características do fenômeno analisado, proporcionando uma compreensão mais profunda e robusta dos dados.

Além disso, aplicamos técnicas de seleção de variáveis, como o método LASSO (Tibshirani, 1996). Das variáveis definidas manualmente, o LASSO selecionou três das quatro variáveis comuns para μ. No entanto, como nosso objetivo é construir um modelo que abranja tanto μ quanto ϕ (pois a análise descritiva sugere que a dispersão não é constante), optamos pelo modelo selecionado utilizando o método descrito nos parágrafos anteriores.

4.1 Considerações sobre o melhor modelo

Na parametrização considerada da distribuição Normal ($NO2$), o parâmetro ϕ representa a variância. A Tabela 4.1 apresenta as estimativas dos parâmetros do modelo final ajustado, bem como estimativas do erro padrão dos estimadores. A tabela traz ainda, para cada parâmetro, o valor da estatística e correspondente valor-p para o teste de hipótese se o parâmetro é ou não igual a zero.

Tabela 4.1: Resultados do Modelo para Média (μ) e Dispersão (ϕ)

Submodelo	**Variável**	**Estimativa**	**Erro Padrão**	**Valor t**	**p-valor**
	(Intercept)	13600,000	520,800	26,105	$< 2 \times 10^{-16}$
	PPC	-0,013	0,00060	-21,075	$< 2 \times 10^{-16}$
μ	IDH	-9876,000	397,000	-24,877	$< 2 \times 10^{-16}$
	TU	-13,480	1,094	-12,323	4×10^{-14}
	TF	-455,400	69,160	-6,585	2×10^{-07}
	(Intercept)	-69,260	6,316	-10,965	1×10^{-12}
	PPC	-0,0007	0,00002	-34,014	$< 2 \times 10^{-16}$
ϕ	IDH	103,300	7,685	13,444	3×10^{-15}
	TU	0,157	0,02121	7,400	1×10^{-08}
	TF	1,401	0,747	1,876	$6,93 \times 10^{-02}$

Como foi usado função de ligação identidade no submodelo para μ, as estimativas dos parâmetros podem ser interpretados como em um modelo linear normal. Assim, temos:

- A estimativa do intercepto (13600) representa o número esperado de mortes por milhão ($\hat{\mu}$) quando todas as covariáveis explicativas ($x_{i1}, x_{i2}, x_{i3}, x_{i4}$) são iguais a zero. Como essas variáveis dificilmente assumem o valor zero, o intercepto pode não ter uma interpretação direta no contexto prático, mas é necessário para ajustar o modelo.

- x_1: PIB per capita (PPC). Estima-se que para cada aumento de 1 unidade no PIB per capita, espera-se uma redução de 0,013 mortes por milhão, mantendo as outras covariáveis constantes. Embora o impacto pareça pequeno, ele é cumulativo. Estima-se por exemplo, uma redução de 13 mortes por milhão para cada aumento de 1000 dólares no PIB per capita mantidas as demais covariáveis constantes.

- x_2: Índice de Desenvolvimento Humano (IDH). Como o IDH varia entre 0 e 1, para uma melhor interpretação devemos dividir a estimativa obtida por 100. Assim, estima-se que para cada aumento de 0,01 no IDH, espera-se uma redução de cerca de 98,76 mortes por milhão mantendo as outras covariáveis constantes. Isso indica que níveis mais altos de desenvolvimento humano (representados por educação, renda e expectativa de vida) estão fortemente associados à diminuição da mortalidade.

- x_3: Taxa de urbanização (TU em %). Assim, estima-se para cada aumento de 1 ponto percentual na taxa de urbanização, espera-se uma redução de 13,48 mortes por milhão, mantendo as outras covariáveis constantes. Isso reflete o fato de que áreas

urbanas geralmente possuem melhor acesso a serviços de saúde e infraestrutura, embora o impacto seja moderado em relação a outras variáveis.

- x_4: Taxa de fertilidade (TF). Estima-se que, para cada aumento de 1 ponto na taxa de fertilidade (número médio de filhos por mulher ao longo da vida reprodutiva), espera-se uma redução de 455,4 mortes por milhão (mpm), mantendo as demais covariáveis constantes. O sinal da estimativa desse parâmetro foi contrário ao sugerido pela análise descritiva. Porém, na análise inferencial estamos analisando o efeito simultâneo de todas as variáveis presentes no modelo. Assim, é bem razoável imaginar que entre países de mesmo PIB per capita, IDH e taxa de urbanização, países com maior taxa de fertilidade tenham menos mortes, pois isso significa ter uma população mais jovem.

Paro o submodelo da dispersão, como foi usada a função de ligação logarítmica, a interpretação deve ser feita para a exponencial das estimativas. Porém, como estamos bem mais interessados no efeito das covariáveis na média do que na variância, vamos interpretar nesse caso apenas o sinal das estimativas. Como o sinal da estimativa no submodelo da dispersão é negativa para o PIB per capita, há evidências que quanto maior o PIB per capita (PPC), menor é a variância do número de mortes por milhão por COVID-19. Já as demais estimativas são positivas. Assim quanto maior forem as covariáves IDH, TU e TF, maior é a variância do número de mortes por milhão por COVID-19. Contudo, a variável TF não é estatisticamente significativa ($p = 0.0693$) ao nível de significância de 5%, e sua interpretação deve ser feita com cautela.

4.2 Análise de diagnóstico

As suposições de um modelo estatístico são o que garante a validade e a confiabilidade dos resultados. A principal suposição do modelo ajustado neste trabalho é que a distribuição da variável resposta é normal com a média e a variância variando em termos de variáveis preditoras. Outra suposição importante é que a relação entre a média e a variância da variável resposta e cada uma das covariáveis foi estabelecida de forma adequada no modelo. Por fim, supõe-se que o modelo e essas suposições valem para toda a região de valores das covariáveis observadas na amostra. Para verificar essas suposições, podemos utilizar o resíduo quantílico. Se essas suposições tiverem satisfeitas,

o resíduo quantílico terá distribuição normal padrão em amostras pequenas e moderadas e o gráfico dos resíduos pelos valores ajustados deverá apresentar como resultado uma nuvem aleatória de pontos.

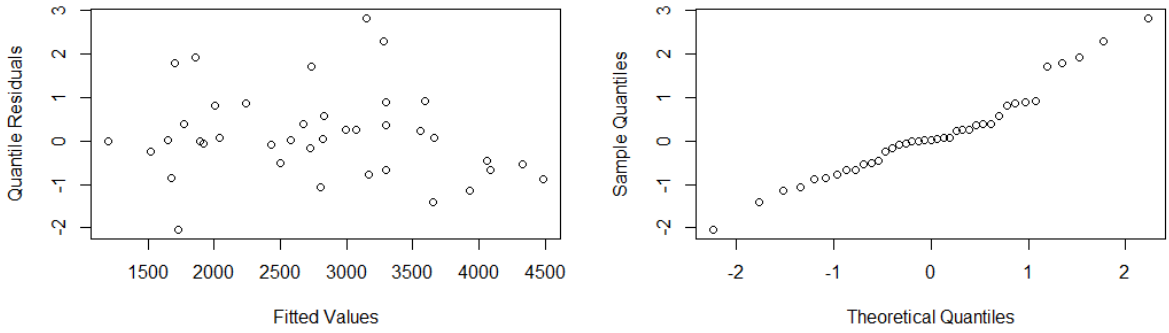

Figura 4.1: Resíduos quantílicos vs valores ajustados para verificar homocedasticidade (esquerda) e Q-Qplot para avaliar normalidade (direita).

A Figura 4.1 apresenta o conjunto de gráficos que o pacote GAMLSS apresenta com o resíduo quantílico. O gráfico esquerdo apresenta os resíduos quantílicos contra os valores ajustados, nota-se a ausência de qualquer padrão sistemático, sugerindo que o modelo é adequado para esses dados. Já o direito, o Q-Q plot, mostra que os resíduos seguem aproximadamente a linha de referência, indicando que a distribuição dos resíduos está próxima da normalidade. Com todas as suposições validadas, podemos concluir que o modelo se mostra adequado para modelar a taxa de mortalidade por COVID-19 nos países europeus, proporcionando uma representação confiável da relação entre a variável resposta mencionada e as covariáveis analisadas.

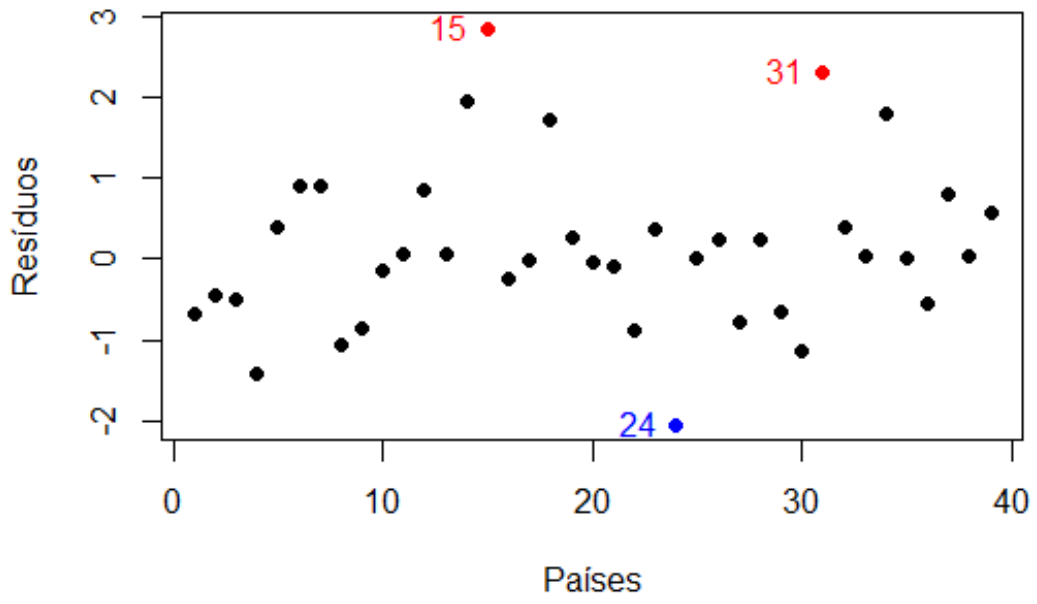

Figura 4.2: Gráfico de valores preditos vs observados

A Figura 4.2 apresenta os resíduos quantílicos, que são utilizados para avaliar a diferença entre os valores observados e os valores esperados com base no modelo ajustado. Esses resíduos permitem verificar quão bem o modelo conseguiu representar o comportamento observado em cada país. Se o modelo estiver bem ajustado, os resíduos quantílicos devem seguir uma distribuição aproximadamente normal. Dessa forma, definimos como pontos atípicos aqueles resíduos que ultrapassam o valor de 2, indicando que o número observado de mortes por COVID-19 foi superior ao previsto pelo modelo (excesso), ou que são inferiores a -2, sugerindo que o número observado foi menor do que o esperado (déficit). Essa definição se justifica pelo fato de que, na distribuição normal padrão, a probabilidade de ocorrer um valor fora do intervalo [-2, 2] é de apenas aproximadamente 5%.

Pela Figura 4.2, os pontos destacados chamam atenção para países que tiveram desvios mais expressivos em relação ao esperado. Os países Hungria (2.84) e Eslováquia (2.32), destacados em vermelho apresentaram os maiores resíduos positivos, indicando que o número de mortes por milhão por COVID-19 nesses países foram significativamente maiores do que o previsto pelo modelo. Por outro lado, temos que a Holanda (-2.05), destacado em azul apresenta o menor resíduo, ou seja, foi o país em que o número de mortes

por milhão por COVID-19 foi bem inferior ao esperado pelo modelo. Com essa análise, foi possível identificar o número de mortes por milhão por COVID-19 nesses países que mais se afasta (positiva ou negativamente) do padrão previsto pelo modelo, auxiliando na compreensão do comportamento dos dados e destacando as nações que mais contribuíram para os extremos da distribuição.

4.3 Discussão sobre os países que se destacaram

Com base nos dados em estudo e na confirmação de que todas as suposições do modelo foram atendidas, estamos interessados em compreender as razões por trás do desempenho desses 3 países na pandemias do COVID-19. Logo, serão levantadas hipóteses relacionadas a vulnerabilidades sociais, desigualdades no acesso à saúde, políticas públicas ineficazes ou eventos específicos que possam ter contribuído para o desfecho.

Essa análise permitirá não apenas destacar os contextos em que houve sucesso ou falhas na resposta a fatores que impactam a mortalidade, mas também levantará questões relevantes para futuras tomadas de decisão. Ao associar os resultados a características específicas de cada país, será possível compreender como as combinações de fatores sociais, políticos e demográficos influenciam o número de mortes, oferecendo insights valiosos para a formulação de políticas públicas mais eficazes.

Países	**Hungria**	**Eslováquia**	**Holanda**
Total de mortes (B)	49048,00	21224,00	22992,00
Mortes por milhão (y)	5106,00	3887,00	1336,00
População	9606259,00	5460193,00	17211447,00
Mortes por milhão esperada ($\hat{\mu}$)	3154,10	3288,77	1736,56
Diferença: $y - \hat{\mu}$	1951,90	598,23	-400,56
Excesso de mortes estimadas (G)	18732,00	3268,00	-6893,00
% de mortes evitáveis (G/B)	38,19%	15,39%	-
% que deixaram de morrer $\left(\frac{\lvert G\rvert}{\lvert G\rvert+B}\right)$	-	-	23,06%

Tabela 4.2: Diferenças entre os países que se destacaram

A Tabela 4.2 apresenta a taxa de mortalidade por COVID-19 (mpm) e algumas outras informações, permitindo estimar o número absoluto de pessoas que morreram a mais ou a menos em relação ao previsto pelo modelo, considerando a população de cada país.

Os resultados revelam variações significativas entre os países em relação ao excesso de mortes, evidenciadas pelas diferenças entre as taxas de mortalidade observadas e as

esperadas. Na Hungria, com uma população de 9.602.659 pessoas, a taxa de mortalidade foi de 5106 (mpm), dado isso, temos que 1951,90 (mpm) foram acima do esperado. Isso corresponde a um excesso estimado de 18.732 mortes, representando 38,19% do total de mortes por COVID-19 registradas no país, sugerindo que uma parcela considerável poderia ter sido evitada.

Na Eslováquia, com uma população de 5.460.193 pessoas, a taxa de mortalidade foi de 3887 (mpm), tendo 598,23 (mpm) acima do esperado. Isso equivale a cerca de 3.268 mortes por COVID-19 a mais do que o previsto, representando 15,39% do total de óbitos, indicando um impacto menos severo em relação à Hungria, mas ainda relevante.

Por outro lado, a Holanda apresentou uma situação distinta. Com uma população de 17.211.447 pessoas, sua taxa de mortalidade foi de 1336 (mpm), apresentando 400,56 (mpm) abaixo do esperado, resultando em 6.893 mortes a menos do que o previsto. Esse déficit corresponde a 23,06% do total estimado de mortes por COVID-19 esperadas, evidenciando que o número de óbitos foi significativamente menor do que o modelo previa, possivelmente refletindo a eficácia de medidas de mitigação.

Esses valores absolutos e percentuais ajudam a contextualizar o impacto real das diferenças entre as taxas de mortalidade observadas e esperadas, revelando disparidades regionais no enfrentamento da pandemia. Enquanto Hungria e Eslováquia enfrentaram excessos substanciais de mortes, a Holanda conseguiu evitar uma proporção significativa de óbitos.

Na Holanda, primeiro-ministro Mark Rutte, líder do Partido Popular para a Liberdade e Democracia de centro-direita, adotou uma abordagem baseada na confiança pública e na responsabilidade individual. Embora inicialmente tenha optado por medidas menos restritivas, a Holanda foi um dos primeiros países da Europa Ocidental a restabelecer as medidas de restrições em menos de 2 meses depois de elas serem flexibilizadas. O governo ajustou rapidamente as políticas em resposta ao aumento de casos, implementando lockdowns e restrições conforme necessário (van Dullemen e Jeanne de Bruijn, 2022). Além disso, apresentou um sistema de saúde robusto e acessível, o que facilitou a testagem em massa e o tratamento eficaz dos pacientes. A alta densidade populacional em áreas urbanas bem conectadas permitiu uma comunicação eficiente das medidas de saúde pública.

Por outro lado, apresentando mortes em excesso, a Hungria foi liderada pelo primeiro-ministro Viktor Orbán, do partido Fidesz, uma agremiação nacional-conservadora de extrema direita. Orbán utilizou a crise sanitária como oportunidade para centralizar poder,

governando por decreto sem limite de tempo definido. Essa abordagem autoritária gerou preocupações internacionais sobre o enfraquecimento da democracia no país. Em termos de gestão da pandemia, a Hungria adotou medidas inicialmente rígidas, mas enfrentou críticas por decisões controversas, como a rápida reabertura da economia, e pela baixa confiança da população em relação às vacinas adquiridas fora da União Europeia, como as da China e da Rússia. Esses fatores, combinados com o envelhecimento populacional e a alta prevalência de comorbidades na população húngara, contribuíram para o elevado número de mortes por milhão. (Welle, 2020)

Na Eslováquia, o primeiro-ministro Igor Matovič, do partido OĽaNO, de centro-direita, enfrentou sérios desafios na gestão da pandemia. Apesar de adotar medidas restritivas iniciais e realizar uma ampla campanha de testes em massa, sua administração foi marcada por controvérsias, como decisões inconsistentes e disputas políticas internas que minaram a eficácia das políticas públicas. A aquisição de vacinas russas Sputnik V, sem aprovação da União Europeia (de Moura, 2021), gerou divisões políticas e desconfiança pública, comprometendo os esforços de vacinação. Além disso, fatores sociais como desigualdades regionais e limitações na infraestrutura de saúde em áreas menos desenvolvidas agravaram o impacto da pandemia no país, contribuindo para o número elevado de mortes.

Um ponto interessante é que a Suécia apresentou um resíduo quantílico de 1,71. Embora não tenha sido um valor acima de 2, isso indica que o país teve um comportamento diferente dos demais países escandinavos, possivelmente devido à estratégia adotada durante a pandemia. Enquanto Noruega e Dinamarca impuseram restrições rigorosas (of Medicine, 2022), a Suécia manteve grande parte das atividades abertas, o que resultou em 2.680 (mpm), um número significativamente superior ao de seus vizinhos. Esse posicionamento explica seu resíduo relativamente alto: os valores reais ficaram acima do esperado, refletindo o impacto de uma política menos restritiva (Brasil, 2020).

Capítulo 5

Considerações finais

O presente estudo analisou a mortalidade por COVID-19 em países europeus utilizando modelos GAMLSS paramétricos (Rigby e Stasinopoulos, 2005), destacando os Modelos Lineares Generalizados Duplos (DGLM) como um caso particular (Smyth, 1989), capaz de modelar simultaneamente a média e a dispersão, considerando a variabilidade não homogênea dos dados. Foram identificadas variáveis socioeconômicas e demográficas, como PIB per capita, IDH, taxa de urbanização e fertilidade, como fatores significativos para explicar as diferenças nas taxas de mortalidade.

Os resultados mostraram que países com maior desenvolvimento humano e melhores condições socioeconômicas apresentaram menores taxas de mortalidade, enquanto nações mais vulneráveis enfrentaram maiores impactos. A abordagem robusta dos DGLM permitiu capturar nuances nos dados, evidenciando a importância de ferramentas avançadas para análises em cenários complexos. Entre as lições aprendidas, destaca-se a relevância da qualidade dos dados, do rigor na seleção de variáveis e da aplicação de modelos estatísticos capazes de lidar com heterogeneidade.

Futuros estudos podem ajustar modelos para o mpm por COVID-19 em outros continentes. Porém, ao contrário da Europa, os demais continentes do planeta apresentam pelo menos alguns países com baixíssimo desenvolvimento. Nesses países, provavelmente a subnotificação de mortes por COVID-19 foi elevada. Assim, para os demais continentes, provavelmente seja mais adequada trabalhar com o excesso de mortes (Organization, 2020) ao invés da mpm por COVID-19. O excesso de mortes refere-se ao número adicional de mortes durante uma crise, acima do esperado em condições normais. Durante a pandemia de COVID-19, ela é uma métrica fundamental para avaliar o impacto total da crise, pois inclui não apenas mortes confirmadas pela doença, mas também aquelas

subnotificadas e decorrentes de outras causas associadas à crise sanitária. Essa medida fornece uma visão mais abrangente dos efeitos da pandemia na taxa de mortalidade. Dessa forma, utilizando-se o excesso de mortes como variável resposta, seria possível desenvolver para outros continentes um estudo semelhante ao desenvolvido neste trabalho sem o risco de obter conclusões equivocadas devido à subnotificações no número de mortes por COVID-19.

Referências Bibliográficas

Alves, J. E. D. (2023). Brasil chega a 700 mil mortes da covid-19. https://www.ecodebate.com.br/2023/03/13/brasil-chega-a-700-mil-mortes-da-covid-19/. Acesso em: 10 Abril 2024.

Amaral, L. S. (2025). Códigos – trabalho de graduação. https://github.com/leojsalles/C-digos-TCC---Modelagem-da-taxa-de-mortes-da-COVID-19-por-pa-s/blob/main/README.md. Acesso em: 20 de Fevereiro 2025.

Bank, T. W. (2024). The world bank. https://www.worldbank.org/en/home. Acesso em: 20 Agosto 2024.

Brasil, B. (2020). Epidemiologista da suécia admite que estratégia gerou mais mortes. https://encurtador.com.br/d1jZk. Acesso em: 03 de Fevereiro de 2025.

Butantan, P. (2021). Vacinas aplicadas em países ricos; número de casos volta a crescer com relaxamento de cuidados. https://tinyurl.com/bda9x57d. Acesso em: 10 Abril 2024.

Canatay, A., Emegwa, T. J. e Talukder, M. F. H. (2021). Critical country-level determinants of death rate during covid-19 pandemic. *International Journal of Disaster Risk Reduction*, **64**, 102507.

Cavalaro, L. L. e Pereira, G. H. (2022). A procedure for variable selection in double generalized linear models. *Journal of Statistical Computation and Simulation*, **92**(13), 2703–2720.

Cole, T. J. e Green, P. J. (1992). Smoothing reference centile curves: the lms method and penalized likelihood. *Statistics in medicine*, **11**(10), 1305–1319.

de Moraes Bernal, H., Siqueira, C. E., Adami, F. e de Sousa Santos, E. F. (2020). Tendência das taxas de letalidade de covid-19 no mundo, entre 2019-2020. *Journal of Human Growth and Development*, **30**(3), 344.

de Moura, I. M. (2021). Eslováquia compra vacina sputnik v e enfrenta controvérsias internas. https://www.gazetadopovo.com.br/mundo/pandemia-derruba-premie-eslovaquia/. Acesso em: 14 de Janeiro de 2025.

Dunn, P. K. e Smyth, G. K. (1996). Randomized quantile residuals. *Journal of Computational and graphical statistics*, **5**(3), 236–244.

Haatveit, B., Westlye, L. T., Vaskinn, A., Flaaten, C. B., Mohn, C., Bjella, T., Sæther, L. S., Sundet, K., Melle, I., Andreassen, O. A. *et al.* (2023). Intra-and inter-individual cognitive variability in schizophrenia and bipolar spectrum disorder: an investigation across multiple cognitive domains. *Schizophrenia*, **9**(1), 89.

in Data, O. W. (2024). Our word in data. https://ourworldindata.org/. Acesso em: 20 Agosto 2024.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. Em *Proceedings of the 10th European Conference on Machine Learning (ECML 1998)*, páginas 137–142. Springer.

Liu, Y., Sinke, L., Jonkman, T. H., Slieker, R. C., Consortium, B., van Zwet, E. W., Daxinger, L. e Heijmans, B. T. (2023). The inactive x chromosome accumulates widespread epigenetic variability with age. *Clinical Epigenetics*, **15**(1), 135.

Maitra, D. S., Bhattacharya, U. e Parui, S. K. (2015). Cnn based common approach to handwritten character recognition of multiple scripts. Em *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, páginas 1021–1025. IEEE.

Montgomery, D. C., Peck, E. A. e Vining, G. G. (2006). *Introduction to linear regression analysis*. John Wiley and Sons, Inc., New York.

MSF (2021). Retrospectiva covid-19. https://www.msf.org.br/noticias/retrospectiva-covid-19-atuacao-de-msf-no-brasil/. Acesso em: 10 Abril 2024.

of Medicine, N. L. (2022). Comparação das taxas de mortalidade por covid-19 na escandinávia. https://encurtador.com.br/0l0yx. Acesso em: 03 de Fevereiro de 2025.

OMS (2021). Folha informativa sobre covid-19. https://www.paho.org/pt/covid19. Acesso em: 10 de Abril 2024.

Organization, W. H. (2020). Excess mortality associated with covid-19 pandemic estimated at 14.9 million in 2020 and 2021. https://encurtador.com.br/RBMPc. Acesso em: 03 de Fevereiro de 2025.

Ornelas, E. (2020). Lockdown 101: Managing economic lockdowns in an epidemic.

Passarelli-Araujo, H., Pott-Junior, H., Susuki, A. M., Olak, A. S., Pescim, R. R., Tomimatsu, M. F., Volce, C. J., Neves, M. A., Silva, F. F., Narciso, S. G. *et al.* (2022). The impact of covid-19 vaccination on case fatality rates in a city in southern brazil. *American journal of infection control*, **50**(5), 491–496.

Paula, G. A. (2013). On diagnostics in double generalized linear models. *Computational Statistics & Data Analysis*, **68**, 44–51.

Pfizer (2022). O que é covid longa e quais os efeitos dela? https://www.pfizer.com.br/noticias/ultimas-noticias/covid-longa. Acesso em: 10 Abril 2024.

R Core Team (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rigby, R. A. e Stasinopoulos, D. (1996). A semi-parametric additive model for variance heterogeneity. *Statistics and Computing*, **6**, 57–65.

Rigby, R. A. e Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society Series C: Applied Statistics*, **54**(3), 507–554.

Rigby, R. A., Stasinopoulos, M. D., Heller, G. Z. e De Bastiani, F. (2019). *Distributions for modeling location, scale, and shape: Using GAMLSS in R*. Chapman and Hall/CRC.

Schölkopf, B. e Smola, A. J. (2002). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT press.

Smyth, G. K. (1989). Generalized linear models with varying dispersion. *Journal of the Royal Statistical Society: Series B (Methodological)*, **51**(1), 47–60.

Sorci, G., Faivre, B. e Morand, S. (2020). Explaining among-country variation in covid-19 case fatality rate. *Scientific reports*, **10**(1), 18909.

Stasinopoulos, D. M. e Rigby, R. A. (2008). Generalized additive models for location scale and shape in (GAMLSS) R. *Journal of Statistical Software*, **23**, 1–46.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **58**(1), 267–288.

Tracker, C. (2024). Banco de dados. https://www.worldometers.info/coronavirus/. Acesso em: 20 agosto 2024.

van Dullemen e Jeanne de Bruijn, C. (2022). Mark rutte and the dutch approach to the pandemic: balancing trust and restrictions. https://ndupress.ndu.edu/Media/News/News-Article-View/Article/2920148/the-dutch-approach-to-covid-19-how-is-it-distinctive/. Acesso em: 14 de Janeiro de 2025.

Vapnik, V. (1995). Support-vector networks. *Machine learning*, **20**, 273–297.

Welle, D. (2020). Hungria não é mais uma democracia, diz relatório. *DW*. Acesso em: 14 de Janeiro de 2025.

Apêndice A

Pacotes utilizados no R

Para a análise e modelagem dos dados neste trabalho, foram empregados os seguintes pacotes no R:

- **ggplot2**: Utilizado para criar visualizações gráficas detalhadas, permitindo explorar a distribuição dos dados e as relações entre variáveis de forma intuitiva.
- **tidyr**: Ferramenta essencial para organização e reestruturação dos dados, facilitando a conversão de tabelas para formatos mais adequados à análise.
- **dplyr**: Conjunto de funções eficientes para manipulação de dados, possibilitando filtragens, agrupamentos e transformações de maneira simplificada e rápida.
- **GGally**: Extensão do `ggplot2` que permite a criação de matrizes de gráficos exploratórios, auxiliando na identificação de padrões e correlações entre variáveis.
- **gamlss**: Pacote fundamental para o ajuste de modelos GAMLSS (Modelos Aditivos Generalizados para Localização, Escala e Forma), possibilitando a modelagem de diferentes distribuições da variável resposta, indo além da média para considerar também aspectos como dispersão e assimetria.

O uso desses pacotes foi crucial para a organização, visualização e modelagem dos dados, garantindo uma abordagem estatística mais robusta e flexível. Os código utilizados na elaboração deste trabalho estão disponíveis no GitHub e podem ser acessados em (Amaral, 2025). Na sequência, é possível visualizar os códigos utilizados para fazer as análises estatísticas deste trabalho.

```r
# LIBERANDO OS PACOTES

library(ggplot2)
library(tidyr)
library(dplyr)
library(GGally)
library(gamlss)

# VISUALIZANDO O BANCO DE DADOS

View(BASETGB)


#  AN LISE DESCRITIVA

#Selecionando as covariaveis para nossa an lise descritiva
select_dataset<-dplyr::select(BASETGB,
                              Mortes1Milh oDePop,
                              TxMortCard,
                              TaxaMortalidadeInfantil,
                              TaxaFertilidade   ,
                              TaxaPopUrb,
                              IDH,
                              ExpectatVida,
                              PIB,
                              Turismo,
                              Densidade)

# Renomeando as vari veis
select_dataset <- dplyr::rename(select_dataset,
                                Mortes = Mortes1Milh oDePop,
                                TMC = TxMortCard,
                                TMI = TaxaMortalidadeInfantil,
                                TF = TaxaFertilidade,
                                TU = TaxaPopUrb,
                                IDH = IDH,
                                EV = ExpectatVida,
                                PPC = PIB,
                                TUR= Turismo,
```

```
                                DEN=Densidade)



# MATRIZ DE CORRELA  ES
matriz_cor <- cor(select_dataset)
print(matriz_cor)
ggcorr(select_dataset, label=T)
ggcorr(select_dataset, label = TRUE, label_round = 2, label_size = 3,
    label_color = "darkblue", palette = "Blues")

# ESTUDO DA VARI VEL RESPOSTA

# BOXPLOT
p <- ggplot(BASETGB, aes(y = Mortes1Milh oDePop)) +
  geom_boxplot(fill = "darkblue", color = "black") +
  labs(y = "Mortalidade por COVID-19",
       x = "") +
  theme_minimal()

print(p)
summary(BASETGB$Mortes1Milh oDePop)

# HISTOGRAMA
hist(BASETGB$Mortes1Milh oDePop, breaks = 10, freq = TRUE, col = "
    darkblue", main = "",
     xlab = "Mortalidade por COVID-19", ylab = "Frequ ncia")


# COVARI VEIS EM ESTDUO
ggplot(BASETGB, aes(x = TxMortCard, y = Mortes1Milh oDePop)) +
  geom_point(color = "darkblue") +
  labs(
    x = "Taxa de Mortalidade Card aca",
    y = "Mortalidade por COVID-19",
  )

# -------

```

```
78 ggplot(BASETGB, aes(x = TaxaMortalidadeInfantil, y = Mortes1Milh oDePop
       )) +
79   geom_point(color = "darkblue") +
80   labs(
81     x = "Taxa de Mortalidade Infantil",
82     y = "Mortes por 1 Milh o de Popula  o"
83   )
84
85 # -------
86
87 ggplot(BASETGB, aes(x = TaxaFertilidade, y = Mortes1Milh oDePop)) +
88   geom_point(color = "darkblue") +
89   labs(
90     x = "Taxa de fertilidade",
91     y = "Mortes por 1 Milh o de Popula  o"
92   )
93
94 # -------
95
96 ggplot(BASETGB, aes(x = TaxaPopUrb, y = Mortes1Milh oDePop)) +
97   geom_point(color = "darkblue") +
98   labs(
99     x = "Porcentagem da pop. urbana",
100    y = "Mortes por 1 Milh o de Popula  o"
101  )
102
103 # -------
104
105 ggplot(BASETGB, aes(x = IDH, y = Mortes1Milh oDePop)) +
106   geom_point(color = "darkblue") +
107   labs(
108     x = "IDH",
109     y = "Mortes por 1 Milh o de Popula  o"
110  )
111
112 # -------
113
114 ggplot(BASETGB, aes(x = ExpectatVida, y = Mortes1Milh oDePop)) +
115   geom_point(color = "darkblue") +
116   labs(
```

```
117     x = "Esperan a␣de␣vida␣ao␣nascer",
118     y = "Mortes␣por␣1␣Milh o␣de␣Popula  o"
119   )
120
121 # -------
122
123 ggplot(BASETGB, aes(x = PIB, y = Mortes1Milh oDePop)) +
124   geom_point(color = "darkblue") +
125   labs(
126     x = "PIB␣Per␣Capita",
127     y = "Mortes␣por␣1␣Milh o␣de␣Popula  o"
128   )
129
130 # -------
131
132 ggplot(BASETGB, aes(x = Turismo, y = Mortes1Milh oDePop)) +
133   geom_point(color = "darkblue") +
134   labs(
135     x = "Taxa␣de␣turistas",
136     y = "Mortes␣por␣1␣Milh o␣de␣Popula  o"
137   )
138
139 # -------
140
141 View(BASETGB)
142 ggplot(BASETGB, aes(x = Densidade, y = Mortes1Milh oDePop)) +
143   geom_point(color = "darkblue") +
144   labs(
145     x = "Densidade␣populacional",
146     y = "Mortalidade␣por␣COVID-19",
147   )
148
149
150
151 # CONSTRUINDO O MODELO
152
153
154
155 modelo_NO2   <- gamlss( formula = BASETGB$Mortes1Milh oDePop ~
156                                     BASETGB$PIB + BASETGB$IDH +
```

```
                                        BASETGB$TaxaPopUrb + BASETGB$
                                        TaxaFertilidade,

                                    sigma.formula=~BASETGB$PIB + BASETGB
                                        $IDH + BASETGB$TaxaPopUrb +
                                        BASETGB$TaxaFertilidade,
                                    family = NO2,
                                    data = na.omit(BASETGB),
                                    trace = FALSE
)
summary(modelo_NO2)
plot(modelo_NO2)
wp(modelo_NO2, ylim.all = 3)
residuos_modelo_NO2 <- residuals(modelo_NO2)
shapiro.test(residuos_modelo_NO2)
ks.test(residuos_modelo_NO2, "pnorm", mean = mean(residuos_modelo_NO2),
    sd = sd(residuos_modelo_NO2))



#  ANALISANDO OS RES DUOS DOS PA SES


plot(residuos_modelo_NO2_IDH_FERT111)
residuos_modelo_NO2_IDH_FERT111
sort(residuos_modelo_NO2_IDH_FERT111)
fitted(modelo_NO2_IDH_FERT111)


# CRIANDO O GR FICO

par(mar = c(5, 4, 4, 2) - 0.1)
plot(residuos_modelo_NO2_IDH_FERT111,
     ylab = "Res duos",
     xlab = "Pa ses",
     pch = 16, col = "black")


indices_menores <- order(residuos_modelo_NO2)[1]
indices_maiores <- order(residuos_modelo_NO2, decreasing = TRUE)[1:2]
```

```
192
193
194 points(indices_menores,
195        residuos_modelo_NO2_IDH_FERT111[indices_menores],
196        col = "blue",
197        pch = 19)
198
199
200 points(indices_maiores,
201        residuos_modelo_NO2_IDH_FERT111[indices_maiores],
202        col = "red",
203        pch = 19)
204
205
206 text(indices_menores,
207      residuos_modelo_NO21[indices_menores],
208      labels = indices_menores,
209      pos = 2, col = "blue")
210
211
212 text(indices_maiores,
213      residuos_modelo_NO2[indices_maiores],
214      labels = indices_maiores,
215      pos = 2, col = "red")
```