Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Estatística

Dissertação de Mestrado

Regressão Beta Ampliada

por

Talia Alves Xavier

Brasília, 28 de janeiro de 2026

Regressão Beta Ampliada

por

Talia Alves Xavier

Dissertação apresentada ao Departamento de Estatística da Universidade de Brasília, como parte dos requisitos necessários para obtenção do título de Mestre em Estatística.

Orientadora: Profª. Terezinha Késsia de Assis Ribeiro

Brasília, 28 de janeiro de 2026

Dissertação submetida ao Programa de Pós-Graduação em Estatística do Departamento de Estatística da Universidade de Brasília como parte dos requisitos para a obtenção do título de Mestre em Estatística.

Texto aprovado por:

Profª. Terezinha Késsia de Assis Ribeiro
Orientadora, EST/UnB

Profª. Cira Etheowalda Guevara Otiniano
EST/UnB

Profª. Jeniffer Johana Duarte Sanchez
DEMA/UFC

Agradecimentos

Essa nova fase foi repleta de desafios, para além da academia, principalmente no que se refere a conciliar o tempo entre estudos, carreira profissional e família. Por esse motivo, eu gostaria de agradecer a todos que de alguma forma contribuíram para que eu me mantivesse focada e concluísse essa etapa tão importante, desde uma palavra encorajadora até o compartilhamento de conhecimentos.

Sou grata, em primeiro lugar, a minha família por todo carinho, apoio, incentivo e compreensão com as ausências durante essa jornada. Eu os amo e sempre serão a minha referência de humanidade e caráter, assim como a minha inspiração para seguir em frente. Espero sempre honrá-los.

Agradeço imensamente a minha orientadora, Terezinha Ribeiro, por toda sua contribuição, dedicação, amizade e pela confiança depositada em mim. Além desta orientação, tive o privilégio de tê-la como professora e não poderia deixar de destacar o quanto a admiro por ser uma profissional brilhante, competente e comprometida. Obrigada pelo apoio e pelos conselhos, que não se limitaram a este trabalho, mas também me ajudaram a pensar em novos passos para minha trajetória.

Agradeço aos meus queridos amigos por me incentivarem e me animarem a enfrentar esse desafio. Em especial, Thiago e Alisson, que me encorajaram a avançar na carreira acadêmica e que me mostraram ser possível transformar o conhecimento em retorno social, através da pesquisa científica, conduzida com ética, seriedade e comprometimento.

Também agradeço aos colegas de curso pelas trocas de aprendizado e experiências, espe-

cialmente ao Matheus, que se tornou um amigo e com quem pude contar durante toda essa jornada. Estendo os agradecimentos aos professores do curso de pós-graduação em estatística da Universidade de Brasília, em particular à Cira Etheowalda.

Por fim, agradeço à agência de fomento. O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

Resumo

Regressão Beta Ampliada

A distribuição beta é frequentemente utilizada para modelar dados contínuos limitados ao intervalo unitário $(0, 1)$, como taxas e proporções, e isto se deve ao fato de ser um modelo flexível capaz de acomodar diferentes formas, incluindo formatos unimodais, de J e J invertido, e de U. No entanto, sua aplicação em situações em que os dados apresentam bimodalidade é limitada, pois a distribuição beta não acomoda formas bimodais. Sendo assim, a presente dissertação de mestrado propõe uma nova distribuição de probabilidade denominada de beta ampliada com o objetivo de acomodar a modelagem de fenômenos aleatórios bimodais que permitam interpretação simples. Propõe-se uma nova classe de modelos de regressão baseada na nova distribuição. A estrutura de regressão é construída de tal forma que se garanta interpretação simples da relação entre resposta e covariáveis, no tocante à mediana da variável resposta. A inferência dos parâmetros sob o novo modelo foi desenvolvida com base no método de máxima verossimilhança. O novo modelo foi implementado utilizando a estrutura dos Modelos Aditivos Generalizados para Localização, Escala e Forma - GAMLSS, via pacote `gamlss` do software `R`. Por fim, ilustrou-se a aplicabilidade do novo modelo de regressão através de dados climáticos reais, nos quais a umidade relativa do ar foi modelada e verificou-se que o novo modelo de regressão apresentou um ajuste aos dados superior ao tradicional modelo de regressão beta.

Palavras-chave: Bimodalidade; Distribuição beta ampliada; Estimador de máxima verossimilhança; GAMLSS; Regressão beta ampliada; Umidade relativa do ar.

Abstract

Extended Beta Regression

The beta distribution is frequently used for continuous data restricted to the unit interval (0, 1), such as rates and proportions, due to its high flexibility in accommodating different shapes, including unimodal, J-shaped, reverse J-shaped, and U-shaped forms. However, its application is limited in situations where the data exhibit bimodality, since the beta distribution does not accommodate bimodal shapes. Therefore, this master's dissertation proposes a new probability distribution, called the extended beta distribution, with the aim of accommodating the modeling of bimodal random phenomena in a way that allows for simple interpretation. A new class of regression models based on this new distribution is proposed. The regression structure is constructed in such a way as to ensure a simple interpretation of the relationship between the response and the covariates, with regard to the median of the response variable. Parameter inference under the new model was developed using the maximum likelihood method. The new model was implemented using the Generalized Additive Models for Location, Scale, and Shape (GAMLSS) framework, through the `gamlss` package of the `R` software. Finally, the applicability of the new regression model was illustrated using real climatic data, in which relative air humidity was modeled, and it was found that the new regression model showed a better fit to the data than the traditional beta regression model.

Keywords: Bimodality; Extended beta distribution; Maximum likelihood estimator; GAMLSS; Extended beta regression; Relative Humidity.

Sumário

Lista de Tabelas

Lista de Figuras

Capítulo 1

Introdução

1.1 Considerações iniciais

A distribuição beta é um modelo probabilístico associado a variáveis aleatórias contínuas definidas no intervalo contínuo unitário $(0, 1)$, sendo particularmente útil para modelar taxas e proporções (Gupta e Nadarajah, 2004). Este modelo se mostra bastante flexível, pois a combinação de seus parâmetros permite acomodar diversos comportamentos da distribuição, tais como formas unimodais, J, J invertido e U (Ferrari e Cribari-Neto, 2004). Tal característica favorece a análise de diferentes fenômenos aleatórios por meio da distribuição beta.

A aplicabilidade da distribuição beta pode ser encontrada em diversas áreas do conhecimento desde estudos clássicos até desenvolvimentos recentes. A título de exemplo, Ji et al. (2005) dividiram as correlações de níveis de expressão gênica em várias populações por meio de um modelo de mistura de distribuições beta. Leite e Virgens Filho (2013) ajustaram adequadamente a velocidade média de ventos através da distribuição beta. Schmidt, Moraes e Migon (2015) utilizaram um modelo de regressão beta para estudar o desempenho padronizado das escolas nas Olimpíadas Brasileiras de Matemática para Escolas Públicas, entre 2006 e 2013. Bernardo, Almeida e Nascimento (2020) utilizaram um modelo de regressão beta inflacionada para o Índice de Qualidade Geral da Educação Municipal (IQGEM), considerando como variá-

veis explicativas o orçamento educacional e fatores sociais. Ribeiro e Ferrari (2023) modelaram a proporção de atum tropical pescado em função da temperatura da superfície do mar por meio de um modelo de regressão beta robusto. Ainda, Majumdar et al. (2024) propuseram uma nova família de misturas de distribuição beta para modelar a proporção de metilação do DNA, que permite identificar limites objetivos para classificação dos estados de metilação em hipometilados, intermediários e hipermetilados, bem como identificar sítios CpG com graus de metilação distintos entre amostras benignas e tumorais.

No tocante a modelos estatísticos que se baseiam na distribuição beta, destaca-se o trabalho de Ferrari e Cribari-Neto (2004) por meio do desenvolvimento dos modelos de regressão beta, permitindo estabelecer relações entre uma variável dependente (resposta) e variáveis independentes (covariáveis) por meio de uma estrutura de regressão para modelar a média da distribuição beta. Nesse contexto, Smithson e Verkuilen (2006) estenderam o modelo de regressão beta, propondo uma estrutura de regressão linear para o parâmetro de precisão ϕ, permitindo, portanto, sua modelagem em função das covariáveis. Também evidencia-se o trabalho de Simas, Barreto-Souza e Rocha (2010), no qual propuseram uma extensão dos modelos de regressão beta de Ferrari e Cribari-Neto (2004) e Smithson e Verkuilen (2006), permitindo estruturas de regressão não lineares para ambos os parâmetros de média e de precisão.

Observa-se, porém, que, apesar da flexibilidade da distribuição beta, sua aplicação direta para modelar fenômenos bimodais ainda é incipiente na literatura, uma vez que a distribuição beta não comporta bimodalidade. O comportamento bimodal de dados pode ser originado por diferentes fatores, dentre eles Vila e Niyazi Çankaya (2021) destacam que podem advir de características próprias dos fenômenos; erros de medição; e questões no delineamento experimental. Nesse sentido, torna-se importante estudar e desenvolver formas de analisar fenômenos que possuem essa característica por natureza.

Para lidar com a modelagem de dados bimodais que variam no intervalo $(0, 1)$ diversos trabalhos têm recorrido a modelos de misturas de distribuições beta, devido a sua flexibilidade para capturar assimetria, heterogeneidade e multimodalidade. A esse respeito, destacam-se Heis-

terkamp e Pennings (2004), que aplicaram misturas finitas de betas à análise de microarrays relacionados a genes, e Migliorati, Di Brisco e Ongaro (2018), que propuseram uma nova distribuição de probabilidades denominada de beta flexível a partir da mistura de duas distribuições beta com médias diferentes, que, segundo os autores, é capaz de lidar com caudas pesadas, assimetria e bimodalidade. Contudo, o uso de modelos de mistura usualmente está associado com a adição de muitos parâmetros ao modelo, o que exige maior trabalho computacional e pode gerar problemas de identificabilidade. Nesse contexto, Schröder e Rahmann (2017) destacam que, em se tratando de distribuições beta, a estimação dos parâmetros por máxima verossimilhança é prejudicada pela singularidade do logaritmo da função de verossimilhança quando algumas observações assumem valores 0 ou 1. Quanto a misturas de distribuições beta, Schröder e Rahmann (2017) destacam, ainda, os seguintes problemas do algoritmo *Expectation-Maximization* (EM): os pesos da mistura não estão bem definidos e a etapa de maximização não pode ser realizada para observações iguais a 0 ou 1; as misturas são sensíveis a perturbações nos dados; e, elevada carga computacional para inúmeras iterações.

Com relação aos modelos para dados contínuos unitários que acomodam bimodalidade, Vila et al. (2024) propuseram uma distribuição beta bimodal, sob a qual propuseram uma estrutura de regressão. No entanto, os parâmetros associados à distribuição proposta não possuem interpretação simples e, portanto, o trabalho não inclui a interpretação clara para os coeficientes da regressão envolvidos, o que dificulta a compreensão dos efeitos das variáveis explicativas na resposta, limitando sua aplicabilidade prática.

Nesse contexto, com o intuito de propor uma alternativa ao modelo de regressão beta que comporte bimodalidade e possua interpretação simples, nesta dissertação propõe-se o desenvolvimento de uma nova distribuição de probabilidades denominada de beta ampliada (BA), na qual ao se aplicar a reparametrização de Ferrari e Cribari-Neto (2004), obtém-se a distribuição beta ampliada reparametrizada (BAR). Para construir uma nova distribuição de probabilidades que comporte bimodalidade, partiu-se da transformação $T(x) = (x - m)|x - m|^{\delta}$ avaliada na função de distribuição acumulada da variável aleatória $Y = X + c$, em que X segue uma

distribuição beta e c uma constante. O método utilizado aqui para gerar a nova distribuição de probabilidades se assemelha às ideias propostas por Otiniano et al. (2023) e Silva, Otiniano e Nakano (2024) que estendem a distribuição generalizada de valores extremos e a distribuição Weibull, respectivamente, de tal modo que comportassem bimodalidade.

A nova distribuição generaliza a distribuição beta tradicional e acomoda a modelagem de fenômenos que possuam duas modas em um determinado intervalo contínuo. Ressalta-se que a distribuição também acomoda dados unimodais e o suporte pode ir além do intervalo (0,1), acomodando, inclusive, os valores 0 e 1. Adicionalmente, desenvolve-se uma estrutura de regressão sob esta nova distribuição e, consequentemente, propõe-se a inferência dos parâmetros, por meio do método de máxima verossimilhança. Uma importante característica do modelo proposto aqui consiste no fato de que a interpretação dos parâmetros de regressão permite mapear as relações entre variável resposta e covariáveis de forma simples, em termos de variações na mediana da distribuição beta ampliada.

Para viabilizar o ajuste do novo modelo, a implementação foi realizada utilizando a estrutura dos Modelos Aditivos Generalizados para Localização, Escala e Forma - GAMLSS, proposta por Rigby e Stasinopoulos (2005), por meio do pacote `gamlss` disponível no software `R` (Rigby e Stasinopoulos, 2005), que oferece uma interface amigável para implementação de novas distribuições. Para avaliar a qualidade do ajuste sob o novo modelo de regressão, propõe-se a utilização dos resíduos quantílicos, cuja vantagem reside no fato de estes resíduos seguirem uma distribuição normal padrão, independentemente da distribuição da variável resposta, desde que o modelo postulado esteja especificado corretamente (Dunn e Smyth, 1996).

Para ilustrar a utilidade da distribuição BAR, bem como do modelo de regressão associado, é apresentada uma aplicação a dados climáticos reais do estado de Goiás, sendo a umidade relativa mínima do ar (URA) a variável resposta estudada. O primeiro resultado observado refere-se ao ajuste considerando os dados como independentes e identicamente distribuídos da variável URA. O segundo ajuste considera covariáveis disponíveis.

Esta dissertação está dividida em cinco capítulos, sendo o primeiro deles correspondente a

esta introdução e aos objetivos do trabalho. No Capítulo 2 são apresentadas as distribuições beta tradicional, beta deslocada, beta bimodal proposta por Vila et al. (2024) e misturas de distribuições beta. No Capítulo 3 é introduzida a generalização da distribuição beta, a distribuição beta ampliada, proposta neste trabalho. No Capítulo 4 são apresentadas as estruturas de regressão baseadas na nova distribuição e a respectiva inferência com base no método de máxima verossimilhança. No Capítulo 5 é realizada uma aplicação a dados reais dos modelos propostos e dos modelos de regressão beta tradicional, incluindo a comparação dos ajustes de ambos. Por fim, o Capítulo 6 compreende as considerações finais.

1.2 Objetivos

1.2.1 Objetivo geral

Propor uma nova distribuição com base na distribuição beta, que acomode dados que apresentam bimodalidade, e desenvolver uma estrutura de regressão a partir da nova distribuição.

1.2.2 Objetivos específicos

- Desenvolver uma nova distribuição baseada no modelo beta que acomode bimodalidade;
- identificar como os parâmetros da nova distribuição podem ser interpretados;
- definir uma classe de modelos de regressão sob a nova distribuição;
- desenvolver o procedimento inferencial para os parâmetros do modelo com base no método de máxima verossimilhança;
- realizar a implementação do modelo por meio do pacote `gamlss` do software `R`;
- aplicar o modelo proposto a dados reais;
- disponibilizar um repositório github com os códigos.

Capítulo 2

Modelos Beta

Este capítulo apresenta uma explanação das distribuições beta tradicional e beta deslocada, bem como destaca o modelo de regressão associado à distribuição beta. Essas distribuições serão essenciais para a compreensão do desenvolvimento da nova distribuição de probabilidades denominada de beta ampliada, no capítulo seguinte. No que se refere a outras abordagens para lidar com a bimodalidade, serão brevemente mencionadas a distribuição beta bimodal proposta por Vila et al. (2024) e a mistura de distribuições beta. Entretanto, essas metodologias não serão empregadas neste trabalho, uma vez que o propósito consiste no desenvolvimento de um modelo de regressão com fácil interpretação, restringindo-se, portanto, à distribuição beta e distribuição beta ampliada reparametrizada (BAR).

2.1 Distribuição beta com dois parâmetros

Considerando o objetivo de obter uma nova distribuição de probabilidades baseada na distribuição beta, inicia-se pela definição: uma variável aleatória contínua X tem distribuição beta com parâmetros $a > 0$ e $b > 0$, se sua função de densidade de probabilidade (FDP) e sua função de distribuição acumulada (FDA) são dadas, respectivamente, por

$$f(x; a, b) = \frac{1}{B(a, b)} x^{a-1}(1 - x)^{b-1}, \quad x \in (0, 1), \tag{2.1}$$

$$F(x; a, b) = \frac{B_x(a, b)}{B(a, b)}, \tag{2.2}$$

em que $B(a, b)$ denota a função beta definida por

$$B(a, b) = \int_0^1 t^{a-1}(1-t)^{b-1}dt = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}, \quad a > 0,\ b > 0,$$

em que $B_x(a, b)$ denota a beta incompleta, dada por

$$B_x(a, b) = \int_0^x z^{a-1}(1-z)^{b-1}dz,$$

com $\Gamma(\cdot)$ denotando a função gama, definida por

$$\Gamma(z) = \int_0^\infty u^{z-1}\mathrm{e}^{-u}du, z > 0.$$

Denota-se por $X \sim \mathrm{B}(a, b)$, quando X segue uma distribuição beta de parâmetros a e b. O k-ésimo momento de X é dado por

$$\begin{aligned} \mathrm{E}(X^k) &= \frac{1}{B(a, b)} \int_0^1 x^{(a+k)-1}(1-x)^{b-1}\,dx \\ &= \frac{1}{B(a, b)} \frac{\Gamma(a+k)\Gamma(b)}{\Gamma(a+b+k)} \\ &= \frac{\Gamma(a+b)}{\Gamma(a)} \frac{\Gamma(a+k)}{\Gamma(a+b+k)}. \end{aligned} \tag{2.3}$$

O primeiro e o segundo momento de $X \sim \mathrm{B}(a, b)$ são obtidos ao substituir k em (2.3) por 1 e 2, respectivamente. Segue que a esperança e a variância de X são dadas, respectivamente, por

$$\mathrm{E}(X) \quad = \quad \frac{\Gamma(a+b)}{\Gamma(a)} \frac{\Gamma(a+1)}{\Gamma(a+b+1)} = \frac{a}{(a+b)},$$

$$\begin{aligned}\mathrm{Var}(X) &= \mathrm{E}(X^2) - \mathrm{E}^2(X) \\ &= \frac{ab}{(a+b)^2(a+b+1)}.\end{aligned}$$

Convém destacar a existência de reparametrizações da distribuição beta que facilitam a interpretação dos coeficientes no contexto de modelos de regressão. A esse respeito, destaca-se o trabalho de Ferrari e Cribari-Neto (2004), no qual consideram a reparametrização da distribuição beta, de modo que $\mu = a/(a+b)$ representa a média da distribuição, enquanto $\phi = a + b$ é um parâmetro de precisão. Portanto, tem-se que

$$a = \mu\phi \quad \text{e} \quad b = (1-\mu)\phi. \tag{2.4}$$

A título de exemplo, pode-se ainda mencionar a reparametrização proposta por Rigby et al. (2019), a qual considera a média $\mu = a/(a+b)$ e um parâmetro de escala $\sigma = (a+b+1)^{-1/2}$ da distribuição, com $0 < \mu < 1$ e $0 < \sigma < 1$. Isso implica em $a = \mu(1-\sigma^2)/\sigma^2$ e $b = (1-\mu)(1-\sigma^2)/\sigma^2$. Essa reparametrização será objeto de estudo no Capítulo 5, referente à aplicação desenvolvida neste trabalho.

A reparametrização proposta por Ferrari e Cribari-Neto (2004) será utilizada para a obtenção da nova distribuição proposta neste trabalho. Deste modo, ao substituir (2.4) em (2.1) e (2.2), obtém-se a FDP e FDA da distribuição beta reparametrizada dadas, respectivamente, por

$$f(x;\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} x^{\mu\phi-1}(1-x)^{(1-\mu)\phi-1}, \quad x \in (0,1),$$

$$F(x;\mu,\phi) = \frac{B_x(\mu\phi,(1-\mu)\phi)}{B(\mu\phi,(1-\mu)\phi)}.$$

Denota-se por $X \sim \text{Beta}(\mu, \phi)$. Neste caso, tem-se

$$\text{E}(X) = \frac{a}{a+b} = \frac{\mu\phi}{\phi} = \mu,$$

$$\begin{aligned} \text{Var}(X) &= \frac{\mu\phi(1-\mu)\phi}{\{\mu\phi + [(1-\mu)\phi]\}^2\{\mu\phi + [(1-\mu)\phi] + 1)\}} \\ &= \frac{\mu(1-\mu)}{(\phi+1)}. \end{aligned}$$

2.2 Distribuição beta com três parâmetros

Seja X uma variável aleatória beta com FDP (2.1); $X \sim \text{B}(a, b)$, e seja, por conveniência, $c \in (-1, 0]$ uma constante. Então, a variável aleatória $Y = X + c$ segue uma distribuição beta "deslocada" (BD) com parâmetros $a > 0$, $b > 0$, e parâmetro de localização c com FDP e FDA de Y dadas, respectivamente, por

$$\begin{aligned} f_Y(y; a, b, c) &= f_X(y - c; a, b) \\ &= \frac{1}{B(a,b)}(y-c)^{a-1}(1-y+c)^{b-1}, \quad y \in (c, 1+c), \end{aligned} \tag{2.5}$$

$$\begin{aligned} F_Y(y; a, b, c) &= F_X(y - c; a, b) \\ &= \frac{B_{y-c}(a,b)}{B(a,b)} \\ &= \frac{1}{B(a,b)} \int_0^{y-c} z^{a-1}(1-z)^{b-1} dz. \end{aligned} \tag{2.6}$$

Note que o x na distribuição beta original, de parâmetros a e b, apenas foi substituído por $x = y - c$. Denota-se por $Y \sim \text{BD}(\text{a}, \text{b}, \text{c})$.

A função quantílica de $X \sim \mathrm{BD(a, b, c)}$ é obtida a partir de

$$\begin{aligned} Q(p; a, b, c) &= F_X^{-1}(y - c; a, b) \\ &= c + F^{-1}(p; a, b), \end{aligned} \tag{2.7}$$

em que $0 < p < 1$.

Considerando a reparametrização da distribuição beta de Ferrari e Cribari-Neto (2004) (em 2.4) e aplicando-a em (2.5) e (2.6), a FDP e FDA da distribuição BD tornam-se, respectivamente,

$$\begin{aligned} f_Y(y; \mu, \phi, c) &= f_X(y - c; \mu, \phi) \\ &= \frac{1}{B(\mu\phi, (1-\mu)\phi)}(y-c)^{\mu\phi-1}(1-y+c)^{(1-\mu)\phi-1}, \end{aligned}$$

com $y \in (c, 1 + c)$, e

$$\begin{aligned} F_Y(y; \mu, \phi, c) &= F_X(y - c; \mu, \phi) \\ &= \frac{B_{y-c}(\mu\phi, (1-\mu)\phi)}{B(\mu\phi, (1-\mu)\phi)} \\ &= \frac{1}{B(\mu\phi, (1-\mu)\phi)} \int_0^{y-c} z^{\mu\phi-1}(1-z)^{(1-\mu)\phi-1} dz. \end{aligned}$$

2.3 Regressão beta tradicional

Ferrari e Cribari-Neto (2004) introduziram a classe de modelos de regressão beta cuja parametrização do modelo é dada pela média μ_i e precisão ϕ da distribuição. Posteriormente, Smithson e Verkuilen (2006) propuseram uma extensão da regressão beta, contemplando a modelagem tanto da média quanto da precisão da distribuição.

Os modelos de regressão beta a serem considerados neste trabalho são definidos por

(i) $Y_i \overset{\text{ind}}{\sim} \text{Beta}(\mu_i, \phi_i),\ i = 1, \ldots, n;$

(ii) $g_\mu(\mu_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta} \Leftrightarrow \mu_i = g_\mu^{-1}(\boldsymbol{x}_i^\top \boldsymbol{\beta})$,

(iii) $g_\phi(\phi_i) = \boldsymbol{z}_i^\top \boldsymbol{\gamma} \Leftrightarrow \phi_i = g_\phi^{-1}(\boldsymbol{z}_i^\top \boldsymbol{\gamma})$,

em que $\overset{\text{ind}}{\sim}$ denota "se distribuem de forma independente", $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{p_1})^\top \in \mathbb{R}^{p_1}$ e $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{p_2})^\top \in \mathbb{R}^{p_2}$ são vetores de parâmetros de regressão desconhecidos, $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip_1})^\top$ e $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{ip_2})^\top$ são vetores de covariáveis de comprimento p_1 e p_2 ($p_1 + p_2 = p < n$), respectivamente, que são assumidas fixadas e conhecidas. Além disso, $g_\mu(\cdot) : (0,1) \to \mathbb{R}$ e $g_\phi(\cdot) : (0,\infty) \to \mathbb{R}$ são funções de ligação estritamente monótonas e duas vezes diferenciáveis. O modelo introduzido por Ferrari e Cribari-Neto (2004) é um modelo de regressão beta com precisão constante, ou seja, assume-se que $\phi_i = \phi$, para $i = 1, \ldots, n$.

Existem diversas opções para a função de ligação $g_\mu(\cdot)$. As funções de ligação mais utilizadas para modelar parâmetros que assumem valores no intervalo unitário $(0,1)$ são:

- função logit: $g_\mu(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$;
- função probit: $g_\mu(\mu) = \Phi^{-1}(\mu)$;
- função cloglog: $g_\mu(\mu) = \log(-\log(1-\mu))$;
- função cauchit: $g_\mu(\mu) = \tan(\pi(\mu - 0{,}5))$,

em que $\Phi^{-1}(\cdot)$ e $\tan(\cdot)$ denotam a função quantílica da distribuição normal padrão e função tangente, respectivamente.

Em particular, a ligação logit oferece uma interpretação interessante quando comparada as demais ligações. A interpretação de cada coeficiente do submodelo de μ se dá em função da razão de chances. Nesse caso, o aumento da variável explicativa x_j em uma unidade leva a uma variação percentual de e^{β_j} na chance média do evento de interesse ocorrer mantendo as demais covariáveis fixadas.

Para a função de ligação $g_\phi(\cdot)$, as funções mais utilizadas são a ligação logaritmo $g_\phi(\phi) = \log(\phi)$ e a ligação raiz-quadrada $g_\phi(\phi) = \sqrt{\phi}$. Estas são as ligações mais utilizadas quando

modela-se um parâmetro que pertence ao intervalo contínuo $(0, \infty)$. Em particular, a ligação logarítmica oferece uma interpretação mais simples comparada à ligação raiz-quadrada. Ao usar a ligação logarítmica para modelar ϕ tem-se que cada acréscimo de uma unidade na variável explicativa z_j, mantidas as outras fixadas, representa uma variação percentual de e^{γ_j} na precisão da variável resposta.

O procedimento inferencial para os parâmetros desta classe de modelos é usualmente feito com base no estimador de máxima verossimilhança. Seja $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \boldsymbol{\gamma}^\top)^\top$ o vetor de parâmetros que desejamos estimar. O logaritmo da função de verossimilhança sob o modelo de regressão beta é dado por

$$\ell(\boldsymbol{\theta} \mid \boldsymbol{y}) = \sum_{i=1}^{n} \ell_i(\boldsymbol{\theta} \mid \boldsymbol{y}), \tag{2.8}$$

em que $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$ é o vetor observado das n variáveis resposta, e o logaritmo da função de verossimilhança individual é

$$\begin{aligned} \ell_i(\boldsymbol{\theta} \mid \boldsymbol{y}) &= \log\Gamma(\phi_i) - \log\Gamma(\mu_i\phi_i) - \log\Gamma((1-\mu_i)\phi_i) + (\mu_i\phi_i - 1)\log(y_i) \\ &+ [(1-\mu_i)\phi_i - 1]\log(1-y_i). \end{aligned}$$

O estimador obtido pela maximização de (2.8) com relação a $\boldsymbol{\theta}$ é o estimador de máxima verossimilhança para $\boldsymbol{\theta}$, e será denotado por $\widehat{\boldsymbol{\theta}}$.

A estimação dos parâmetros do modelo e demais procedimentos inferenciais podem ser feitos usando a função `betareg` do pacote `betareg` no software `R` (para mais detalhes veja Cribari-Neto e Zeileis (2010)), ou, também, a função `gamlss` do pacote `gamlss` do `R`. Uma diferença entre essas duas abordagens é o fato de que o pacote `betareg` utiliza as reparametrizações de Ferrari e Cribari-Neto (2004) e Smithson e Verkuilen (2006), enquanto o pacote `gamlss` utiliza a reparametrização de Rigby et al. (2019).

2.4 Distribuição beta bimodal

Vila et al. (2024) propuseram a distribuição beta bimodal (Bbeta), que é uma extensão mais flexível da distribuição beta e permite modelar dados contidos no intervalo $(0, 1)$ unimodais e bimodais. A distribuição é obtida a partir de uma transformação quadrática para gerar funções bimodais.

Seja X uma variável aleatória Bbeta com vetor de parâmetros $\boldsymbol{\theta}_\delta = (\alpha, \beta, \rho, \delta)^\top$, com $\alpha > 0$, $\beta > 0$, $\rho \geq 0$ e $\delta \in \mathbb{R}$, denotada por $X \sim \text{Bbeta}(\alpha, \beta, \rho, \delta)$, então a FDP de X é dada por

$$f(x; \boldsymbol{\theta}_\delta) = \frac{\rho + (1 - \delta x)^2}{Z(\boldsymbol{\theta}_\delta) B(\alpha, \beta)} x^{\alpha-1}(1 - x)^{\beta-1}, \quad x \in (0, 1),$$

com

$$Z(\boldsymbol{\theta}_\delta) = 1 + \rho - 2\delta \frac{\alpha}{\alpha + \beta} + \delta^2 \frac{\alpha(\alpha + 1)}{(\alpha + \beta)(\alpha + \beta + 1)},$$

em que α, β e δ são parâmetros de forma e δ controla a existência de bimodalidade, segundo os autores.

A FDA pode ser obtida por

$$F(x; \alpha, \beta, \rho, \delta) = \int_0^x f(t; \alpha, \beta, \rho, \delta)\, dt. \tag{2.9}$$

A distribuição Bbeta apresenta formas analíticas fechadas para os momentos, como média e variância, contudo, essas são funções complexas que dependem dos quatro parâmetros da distribuição. O que impossibilita a reparametrização de tal forma que se desenvolva um modelo de regressão com fácil interpretação dos coeficientes, relacionando as covariáveis com a variável resposta.

2.4.1 Regressão beta bimodal

Vila et al. (2024) propuseram um modelo de regressão associado à distribuição beta bimodal, atribuindo estrutura de regressão aos parâmetros de forma, α e β. Os parâmetros ρ e δ não possuem estrutura de regressão associada, contudo, devem ser estimados conjuntamente. Os modelos de regressão beta bimodal propostos são definidos por

(i) $Y_i \overset{\text{ind}}{\sim} \text{Bbeta}(\alpha_i, \beta_i, \rho, \delta),\ i = 1, \ldots, n;$

(ii) $g_1(\alpha_i) = \eta_{1i} = \boldsymbol{w_i}^\top \boldsymbol{\gamma}$

(iii) $g_2(\beta_i) = \eta_{2i} = \boldsymbol{z}_i^\top \boldsymbol{\zeta},$

com $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)^\top$ e $\boldsymbol{\zeta} = (\zeta_1, \ldots, \zeta_q)^\top$ são vetores desconhecidos dos coeficientes de regressão, que são assumidos independentes, $\boldsymbol{\gamma} \in \mathbb{R}^p$ e $\boldsymbol{\zeta} \in \mathbb{R}^q$ com $p + q < n$, η_{1i} e η_{2i} são os preditores lineares, $w_i = (w_{i1}, \ldots, w_{ip})^\top$ e $z_i = (z_{i1}, \ldots, z_{iq})^\top$ são as observações. A função de ligação proposta para os dois parâmetros é a logarítmica, tendo em vista que $\alpha > 0$ e $\beta > 0$.

Embora o modelo Bbeta possa se ajustar bem a conjuntos de dados que apresentem bimodalidade, possui uma limitação no que se refere a interpretabilidade. Como a estrutura de regressão é atribuída a parâmetros de forma e não sobre um parâmetro de posição, por exemplo, a interpretação dos coeficientes de regressão torna-se complexa, ou seja, os parâmetros não possibilitam mensurar a relação da variável resposta com as covariáveis. Além disso, a implementação do ajuste desse modelo de regressão não está prontamente disponível para reprodução.

2.5 Misturas de distribuições beta

Para definir como se propõe misturas de distribuições, considere o que se segue, conforme descrito em Stasinopoulos et al. (2017). Suponha que a variável aleatória Y provenha de uma população que é particionada em K subpopulações ou componentes, então a FDP de Y é dada por

$$f(y) = \sum_{k=1}^{K} \pi_k f_k(y),$$

onde $k = 1, 2, ..., K$, $0 \leq \pi_k \leq 1$, tal que $\sum_{k=1}^{K} \pi_k = 1$ é a probabilidade de Y vir da k-ésima componente e $f_k(y)$ é FDP de Y na k-ésima componente. A FDA Y é dada por

$$F(y) = \sum_{k=1}^{K} \pi_k F_k(y).$$

Usando o teorema de Bayes, a probabilidade condicional de uma observação pertencer à componente k dado y é

$$p_k = \frac{\pi_k f_k(y)}{f(y)} = \frac{\pi_k f_k(y)}{\sum_{k=1}^{K} \pi_k f_k(y)}.$$

A FDP da k-ésima componente pode depender de parâmetros $\boldsymbol{\theta_k}$ que, por sua vez, podem ser função de variáveis explicativas $\boldsymbol{x_k}$, isto é, $f_k(y) = f_k(y|\boldsymbol{\theta_k})$. Portanto, a FDP de Y depende dos parâmetros $\boldsymbol{\psi} = (\boldsymbol{\theta}, \boldsymbol{\pi})$, $\boldsymbol{\theta} = (\boldsymbol{\theta_1}^\top, \ldots, \boldsymbol{\theta_K}^\top)^\top$, e $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)\top$, com variáveis explicativas $\boldsymbol{x} = (x_1, \ldots, x_K)\top$. Assim,

$$f(y) = f(y|\boldsymbol{\psi}) = \sum_{k=1}^{K} \pi_k f_k(y|\boldsymbol{\theta_k}).$$

No que se refere a misturas de distribuição beta, pode-se citar a distribuição Beta Flexível (BF) proposta por Migliorati, Di Brisco e Ongaro (2018), que possui suporte no intervalo $(0, 1)$ e é definida como uma mistura de duas distribuições beta com mesmo parâmetro de precisão ϕ e médias distintas $\lambda_1 > \lambda_2$. Seja $Y \sim BF(\lambda_1, \lambda_2, \phi, p)$, então sua FDP é dada por

$$f^*_{BF}(y; \lambda_1, \lambda_2, \phi, p) = p f^*_B(y; \lambda_1, \phi) + (1 - p) f^*_B(y; \lambda_2, \phi),$$

em que $0 < \lambda_2 < \lambda_1 < 1$, $\phi > 0$, $0 < p < 1$, e f^*_B é a densidade beta reparametrizada proposta por Ferrari e Cribari-Neto (2004). A esperança e a variância de Y são, respectivamente, dadas por

$$\mathrm{E}(Y) = p\lambda_1 + (1 - p)\lambda_2,$$

$$\text{Var}(Y) = \frac{\text{E}(Y)(1-\text{E}(Y)) + \phi(\lambda_1-\lambda_2)^2 p(1-p)}{\phi+1}.$$

A estrutura de mistura da distribuição BF estende a variedade de formas acomodadas pela distribuição beta, inclusive, em termos de bimodalidade, assimetria e comportamento de cauda, segundo Migliorati, Di Brisco e Ongaro (2018). Também garante que cada componente seja distinguível, evitando o problema de troca de rótulos.

Com intuito de atribuir um modelo de regressão à distribuição BF, Migliorati, Di Brisco e Ongaro (2018) propuseram uma reparametrização, que inclui a média da distribuição, de modo que:

$$\begin{aligned} \mu &= \text{E}(Y) = p\lambda_1 + (1-p)\lambda_2; \\ \phi &= \phi; \\ w &= \frac{\lambda_1-\lambda_2}{\min\left(\frac{\mu}{p}, \frac{1-\mu}{1-p}\right)}; \\ p &= p, \end{aligned} \tag{2.10}$$

em que $\mu \in (0,1)$ é a média de Y, $w \in (0,1)$ é uma medida de distância normalizada entre as duas componentes da mistura, $p \in (0,1)$ é a proporção de mistura e $\phi > 0$ é um parâmetro de precisão.

Os modelos de regressão beta flexível (RBF) considerados, assumindo que Y_i é independentemente distribuído como uma distribuição BF reparametrizada são definidos por

(i) $Y_i \overset{\text{ind}}{\sim} \text{BF}(\mu_i, \phi_i, w, p),\ i = 1, \ldots, n;$

(ii) $g_\mu(\mu_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta} \Leftrightarrow \mu_i = g_\mu^{-1}(\boldsymbol{x}_i^\top \boldsymbol{\beta}),$

(iii) $g_\phi(\phi_i) = \boldsymbol{z}_i^\top \boldsymbol{\delta} \Leftrightarrow \phi_i = g_\phi^{-1}(\boldsymbol{z}_i^\top \boldsymbol{\delta}),$

em que $\boldsymbol{x_i}^\top = (x_{i0}, \ldots, x_{ik})$ e $\boldsymbol{z_i}^\top = (z_{i0}, \ldots, z_{il})$ são vetores de covariáveis; $\boldsymbol{\beta}^\top = (\beta_0, \ldots, \beta_k)$ e $\boldsymbol{\delta}^\top = (\delta_0, \ldots, \delta_l)$, os vetores de parâmetros de regressão; $g_\mu(\cdot)$ e $g_\phi(\cdot)$ são as funções de li-

gação. Considerando o fato de μ_i, w e p pertencerem ao intervalo $(0, 1)$, pode-se empregar a função de ligação logit. Enquanto para $\phi_i > 0$, pode-se empregar a ligação logaritmo.

Segundo os autores, é possível estender os modelos para regressão nos demais parâmetros da distribuição BF. O modelo RBF adota a abordagem Bayesiana, portanto, requer o uso de algoritmos de Monte Carlo via Cadeias de Markov (MCMC), como o amostrador de Gibbs, e diagnósticos de convergência e autocorrelação. Em contrapartida, o modelo de regressão que será proposto neste trabalho baseia-se na máxima verossimilhança, que permite a estimação direta dos parâmetros, simplificando o processo computacional ao evitar etapas como o *burn-in* e o *thinning*. Além disso, ao utilizar essa abordagem paramétrica, não é necessário especificar distribuições a priori para os parâmetros (como normais multivariadas e gamas adotadas no modelo RBF), o que reduz a subjetividade no processo de estimação.

Capítulo 3

Distribuição beta ampliada

Neste capítulo será discutida com detalhes a construção da nova distribuição de probabilidade proposta neste trabalho, denominada de beta ampliada. A distribuição será obtida por meio da composição da distribuição beta deslocada com uma transformação $T(x)$, em que $X \sim \text{BD}(a, b, c)$, resultando na distribuição beta ampliada, que possui 5 parâmetros. Além disso, será explicitada a mediana da nova distribuição, que será essencial para estabelecer a estrutura da classe de modelos de regressão que será proposta neste trabalho. Por fim, será feita uma breve análise gráfica do comportamento da distribuição beta ampliada conforme os parâmetros especificados.

3.1 Distribuição beta ampliada

Para criar a nova distribuição de probabilidades, considere a seguinte transformação:

$$T = T(x) = (x - m)|x - m|^{\delta}, \;\; x \in \mathbb{R}, \; m \in \mathbb{R}, \; \delta > 0, \tag{3.1}$$

que é responsável por gerar bimodalidade, conforme abordado em Silva, Otiniano e Nakano (2024).

Note que a transformação T depende de x, m e δ. Para $x \in (0, 1)$, a derivada de T com

relação ao seu argumento é expressa por

$$\begin{aligned} T'(x) &= \frac{dT(x)}{dx} \\ &= |x-m|^{\delta} + (x-m)\delta|x-m|^{\delta-1}\text{sgn}(x-m) \\ &= |x-m|^{\delta} + \delta|x-m|^{\delta} \\ &= (\delta+1)|x-m|^{\delta}, \end{aligned}$$

em que $\text{sgn}(\cdot)$ denota a função sinal, definida por

$$\text{sgn}(x) = \begin{cases} 1, & \text{se } x > 0 \\ 0, & \text{se } x = 0 \\ -1, & \text{se } x < 0. \end{cases}$$

Essa derivada será necessária para expressar a FDP da nova distribuição mais adiante.

Para a obtenção da função inversa de T, considere:

- se $x - m > 0$, então $T = (x-m)^{\delta+1} > 0$, e

$$T = (x-m)^{\delta+1} \iff T^{1/(\delta+1)} = x - m \iff x = m + T^{1/(\delta+1)};$$

- se $x - m < 0$, então $T = -(m-x)^{\delta+1} < 0$, e

$$T = -(m-x)^{\delta+1} \iff (-T)^{1/(\delta+1)} = m - x \iff x = m - (-T)^{1/(\delta+1)}.$$

Observe que, em ambos os casos, T é uma função monótona crescente em x, e T assume zero quando $x = m$. Portanto, T é uma função monótona crescente, e sua inversa é expressa por

$$T^{-1} = T^{-1}(x) = m + \text{sgn}(x)|x|^{1/(\delta+1)}. \tag{3.2}$$

A inversa da função T é de suma importância para estabelecer o suporte e a função quantílica da nova distribuição de probabilidades proposta a seguir.

Seja $X \sim \mathrm{BD(a,b,c)}$, ao compor a FDA da distribuição beta deslocada (2.6) com a transformação T em (3.1), obtemos uma nova variável aleatória Y com FDA e FDP definidas, respectivamente, por

$$\begin{aligned} F_Y(y;a,b,c,m,\delta) &= F_X(T(y);a,b,c) \\ &= \frac{B_{T(y)-c}(a,b)}{B(a,b)} \\ &= \frac{1}{B(a,b)}\int_0^{T(y)-c} z^{a-1}(1-z)^{b-1}dz \\ &= \frac{1}{B(a,b)}\int_0^{(y-m)|y-m|^{\delta}-c} z^{a-1}(1-z)^{b-1}dz, \end{aligned} \tag{3.3}$$

$$\begin{aligned} f_Y(y;a,b,c,m,\delta) &= f_X(T(y);a,b,c)T'(y) \\ &= \frac{1}{B(a,b)}(T(y)-c)^{a-1}(1-T(y)+c)^{b-1}(\delta+1)|y-m|^{\delta} \\ &= \frac{1}{B(a,b)}\left[(y-m)|y-m|^{\delta}-c\right]^{a-1} \\ &\times \left[1-(y-m)|y-m|^{\delta}+c\right]^{b-1}(\delta+1)|y-m|^{\delta}, \end{aligned} \tag{3.4}$$

dado que $T(y) = (y-m)|y-m|^{\delta} \in (c, 1+c),\ c \in (-1,0]$, $m \in \mathbb{R}$, $\delta > 0$, $a > 0$, e $b > 0$.

Para identificar o suporte da distribuição, considere a transformação T em (3.1), sua inversa T^{-1} em (3.2), e que $T(y) \in (c, 1+c)$. Assim, para que a FDA (3.3) e FDP (3.4) estejam bem definidas, tem-se que

- se $T(y) > c$, então $T^{-1}(y) > T^{-1}(c) = m + \mathrm{sgn}(c)|c|^{1/(\delta+1)}$;
- se $T(y) < 1+c \ \rightarrow\ T^{-1}(y) < T^{-1}(1+c) = m + \mathrm{sgn}(1+c)|1+c|^{1/(\delta+1)}$,

desde que $T(y)$ é uma função monótona crescente em y.

Assim, o suporte da distribuição beta com cinco parâmetros é dado por

$$y \in \left(m + \mathrm{sgn}(c)|c|^{1/(\delta+1)},\ m + \mathrm{sgn}(1+c)|1+c|^{1/(\delta+1)}\right).$$

Desde que $-1 < c < 0$, o suporte se torna:

$$y \in \left(m - |c|^{1/(\delta+1)},\ m + |1+c|^{1/(\delta+1)}\right).$$

Se uma variável aleatória Y possui FDA e FDP expressas por (3.3) e (3.4), respectivamente, denota-se por $Y \sim \mathrm{BA1}(a, b, c, m, \delta)$, e a partir daqui denominaremos esta nova distribuição de distribuição beta ampliada 1 (BA1). Cabe destacar que a distribuição BA1 é definida para variáveis contínuas com suporte que pode assumir valores reais, bem como que a distribuição beta é um caso particular da distribuição BA1 quando fixados os parâmetros $c = 0$, $\delta = 0$ e $m = 0$. Ressalta-se que todo esse desenvolvimento matemático pode ser, e será, realizado de forma análoga com base na reparametrização da distribuição beta em função dos parâmetros μ e ϕ (2.4).

A seguir, será apresentada a função quantílica da distribuição BA1, a qual será utilizada para reparametrizar a distribuição de forma conveniente no contexto de regressão, bem como para gerar números pseudo-aleatórios. Considere o procedimento que se segue.

Seja $\tau_p = Q(p; a, b, c, m, \delta)$, $0 < p < 1$, o quantil de ordem p em que $Q(\cdot; a, b, c, m, \delta)$ denota a função quantílica de $Y \sim \mathrm{BA1}(a, b, c, m, \delta)$. Note que para $p = 0{,}5$, τ_p representa a mediana da distribuição BA1. Como o quantil τ_p satisfaz $p = F_Y(\tau_p; a, b, c, m, \delta)$, segue que

$$\begin{aligned}
\tau_p &= F_Y^{-1}(p; a, b, c, m, \delta) \\
&= T^{-1}(F_X^{-1}(p; a, b, c)) \\
&= T^{-1}(c + F^{-1}(p; a, b)) \\
&= m + \mathrm{sgn}(c + F^{-1}(p; a, b))|c + F^{-1}(p; a, b)|^{1/(\delta+1)},
\end{aligned}$$

em que $0 < p < 1$, e $F^{-1}(\cdot; a, b)$ é a inversa (função quantílica) da FDA da distribuição beta com dois parâmetros, a qual não possui forma analítica fechada, e pode ser obtida numericamente; por exemplo, por meio da função `qbeta` do software `R` (para mais informações

consulte https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/Beta).

Observe que m é uma função de τ_p, conforme segue

$$m = \tau_p - \text{sgn}(c + F^{-1}(p; a, b))|c + F^{-1}(p; a, b)|^{1/(\delta+1)}.$$

Substituindo a expressão de m em (3.3), tem-se a FDA da distribuição beta ampliada 2 (BA2), que depende do quantil de ordem p da distribuição e dos parâmetros a, b, c e δ, ou seja, $X \sim \text{BA2}(a, b, c, \tau_p, \delta)$. O suporte da distribuição BA2 fica expresso por

$$y \in \Big(\tau_p - \text{sgn}(c + F^{-1}(y; a, b))|c + F^{-1}(y; a, b)|^{1/(\delta+1)} - |c|^{1/(\delta+1)},$$
$$|1 + c|^{1/(\delta+1)} + \tau_p - \text{sgn}(c + F^{-1}(y; a, b))|c + F^{-1}(y; a, b)|^{1/(\delta+1)}\Big).$$

Tomando $p = 0{,}5$, denota-se por $\mu_d = \tau_{0,5}$ a mediana da distribuição BA2, e, consequentemente, a distribuição fica parametrizada em termos de um parâmetro que possui interpretação simples. Esta abordagem é de suma importância para a definição da estrutura de regressão sob a distribuição BA2.

3.1.1 Reparametrização da beta ampliada

É possível que a reparametrização (2.4) da distribuição beta apresente maior simplicidade no que diz respeito à interpretabilidade dos parâmetros ou à implementação do modelo de regressão. Considerando essa reparametrização (2.4), obtêm-se as respectivas expressões para a FDA (3.3) e a FDP (3.4) da distribuição BA2 dadas por

$$\begin{aligned} F_Y(y; \mu, \phi, c, \mu_d, \delta) &= F_X(T(y); \mu, \phi, c) \\ &= \frac{1}{B(\mu\phi, (1-\mu)\phi)} \int_0^{(y-m)|y-m|^{\delta}-c} z^{\mu\phi-1}(1-z)^{((1-\mu)\phi)-1} dz, \end{aligned} \quad (3.5)$$

$$\begin{aligned} f_Y(y;\mu,\phi,c,\mu_d,\delta) &= f_X(T(y);\mu,\phi,c)T'(y) \\ &= \frac{1}{B(\mu\phi,(1-\mu)\phi)}(T(y)-c)^{\mu\phi-1}(1-T(y)+c)^{((1-\mu)\phi)-1}(\delta+1)|y-m|^{\delta} \\ &= \frac{1}{B(\mu\phi,(1-\mu)\phi)}\left[(y-m)|y-m|^{\delta}-c\right]^{\mu\phi-1} \\ &\times \left[1-(y-m)|y-m|^{\delta}+c\right]^{((1-\mu)\phi)-1}(\delta+1)|y-m|^{\delta}, \end{aligned} \tag{3.6}$$

em que $T(y) = (y-m)|y-m|^{\delta} \in (c, 1+c),\ c \in (-1,0]$, $m \in \mathbb{R}$, $\delta > 0$, $\mu \in (0,1)$, $\phi > 0$, $\mu_d \in \mathbb{R}$ representa a mediana, m é uma função dos parâmetros que indexam a distribuição, dada por

$$m = \mu_d - \text{sgn}(c + F^{-1}(p;\mu\phi,(1-\mu)\phi))|c + F^{-1}(p;\mu\phi,(1-\mu)\phi)|^{1/(\delta+1)}.$$

Destaca-se que, diferentemente da distribuição beta, aqui o parâmetro μ não se refere à média da distribuição, nem mesmo a alguma característica simples. Tampouco o parâmetro ϕ representa necessariamente a precisão. O suporte da distribuição é dado por

$$\begin{aligned} y \in \Big(&\tau_p - \text{sgn}(c + F^{-1}(y;\mu\phi,(1-\mu)\phi))|c + F^{-1}(y;\mu\phi,(1-\mu)\phi)|^{1/(\delta+1)} - |c|^{1/(\delta+1)}, \\ &|1+c|^{1/(\delta+1)} + \tau_p - \text{sgn}(c + F^{-1}(y;\mu\phi,(1-\mu)\phi))|c + F^{-1}(y;\mu\phi,(1-\mu)\phi)|^{1/(\delta+1)}\Big). \end{aligned}$$

Considere Y com funções de distribuição acumulada (3.5) e densidade de probabilidade (3.6), então Y segue uma distribuição beta ampliada reparametrizada – BAR, denotando-se por $Y \sim \text{BAR}(\mu,\phi,c,\tau_p,\delta)$, especificamente, quando $p = 0{,}5$, $Y \sim \text{BAR}(\mu,\phi,c,\mu_d,\delta)$.

Tendo em consideração a complexidade da distribuição proposta e o tempo disponível para execução, este trabalho não tem como foco calcular outras propriedades como, por exemplo, os momentos da distribuição, como média, variância ou medidas de assimetria e curtose. O foco deste trabalho reside na modelagem da mediana da variável resposta.

3.1.2 Análise gráfica dos parâmetros

Uma estratégia possível para identificação do comportamento de uma função densidade é a análise gráfica desta avaliada em seu suporte e contando com a especificação adequada de seus parâmetros. Nas Figuras 3.1, 3.2, 3.3, 3.4, e 3.5 são apresentadas curvas para a função densidade de probabilidade da distribuição beta ampliada reparametrizada (3.6), geradas a fim de analisar o comportamento dessa distribuição de acordo com a variação de cada parâmetro mantendo os demais fixados. Além disso, cada figura denota o suporte da distribuição, o qual depende dos parâmetros e acomoda dados fora do intervalo $(0, 1)$. As observações sobre o comportamento da densidade da distribuição BAR estão descritas a seguir.

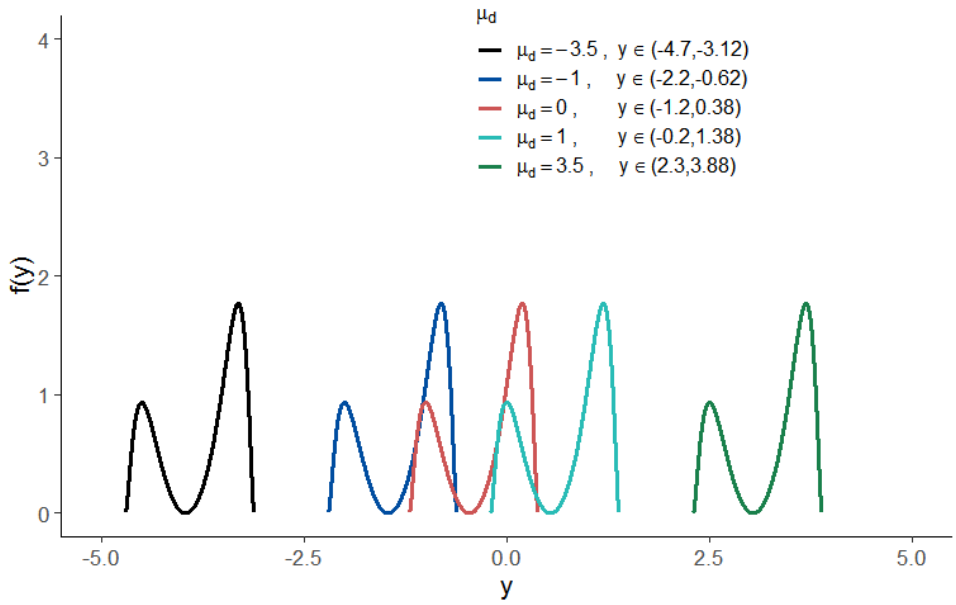
O parâmetro μ_d corresponde à mediana da distribuição BAR e indica o valor de Y para o qual 50% dos valores da distribuição estão abaixo (ou acima). O parâmetro μ_d, como apresentado na Figura 3.1, atua como um parâmetro de localização, indicando a posição dos eventos raros no interior do suporte da distribuição. Por essa razão, de acordo com a variação de μ_d, observa-se o deslocamento das FDPs da distribuição BAR.

Conforme apresentado na Figura 3.2, o parâmetro c afeta a forma da distribuição BAR, e valores de c muito próximos de 0 ou -1 geram distribuições unimodais. Além disso, c parece ter relação com a assimetria da distribuição; na Figura 3.2, $c > -0{,}5$ favorece a assimetria à esquerda e, caso contrário, à direita.

Conforme apresentado na Figura 3.3, a variação de valores de μ causa mudanças consideráveis na curva da distribuição à medida que varia. O efeito que μ parece causar na distribuição beta ampliada reparametrizada se assemelha ao efeito que μ causa na distribuição beta. Na distribuição beta, a moda da distribuição fica próxima a μ, e na distribuição BAR, a maior moda também parece se aproximar de μ.

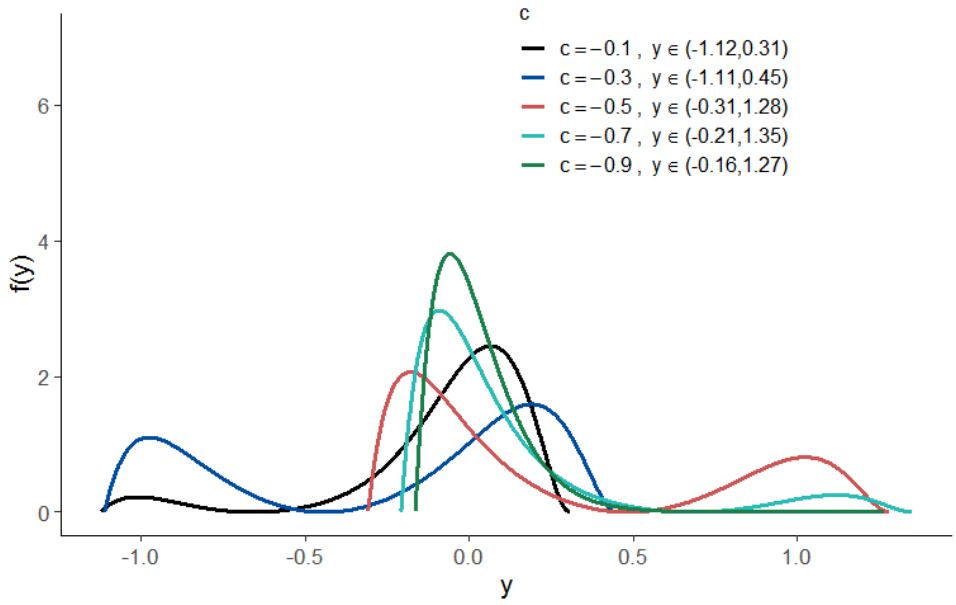
O parâmetro ϕ aparenta estar associado à precisão da distribuição BAR, pois as FDPs dispostas na Figura 3.4 se diferenciam, principalmente, pelo achatamento da curva. Interessante destacar também que, conforme a Figura 3.4, a distribuição beta ampliada reparametrizada aco-

Figura 3.1: Curvas da FDP da distribuição BAR segundo variação dos valores de μ_d, fixados os demais parâmetros ($\mu = 0,5,\ \phi = 5,\ c = -0,4,\ \delta = 2$).



Fonte: Elaboração própria.

Figura 3.2: Curvas da FDP da distribuição BAR segundo variação dos valores de c, fixados os demais parâmetros ($\mu = 0,4,\ \phi = 5,\ \mu_d = 0,\ \delta = 2$).



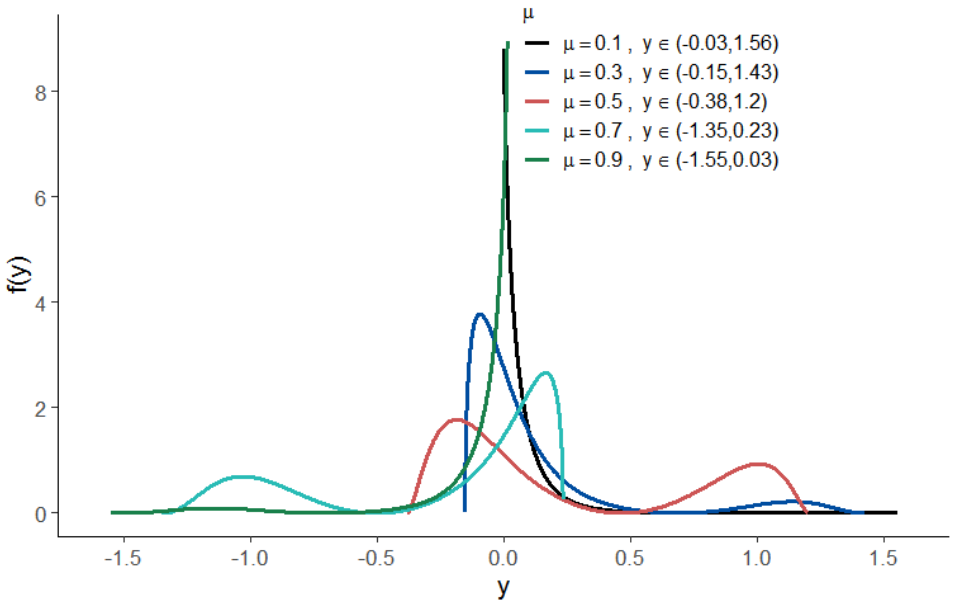
Fonte: Elaboração própria.

moda as formas de U ($\phi = 1$) e de banheira ($\phi = 2$), assim como a distribuição beta usual.

Por fim, quando a função densidade da distribuição beta ampliada apresenta bimodalidade, o parâmetro δ parece afetar a distância entre as modas da distribuição, de forma que quanto maior o valor de δ, maior o afastamento entre elas, conforme apresentado na Figura 3.5.

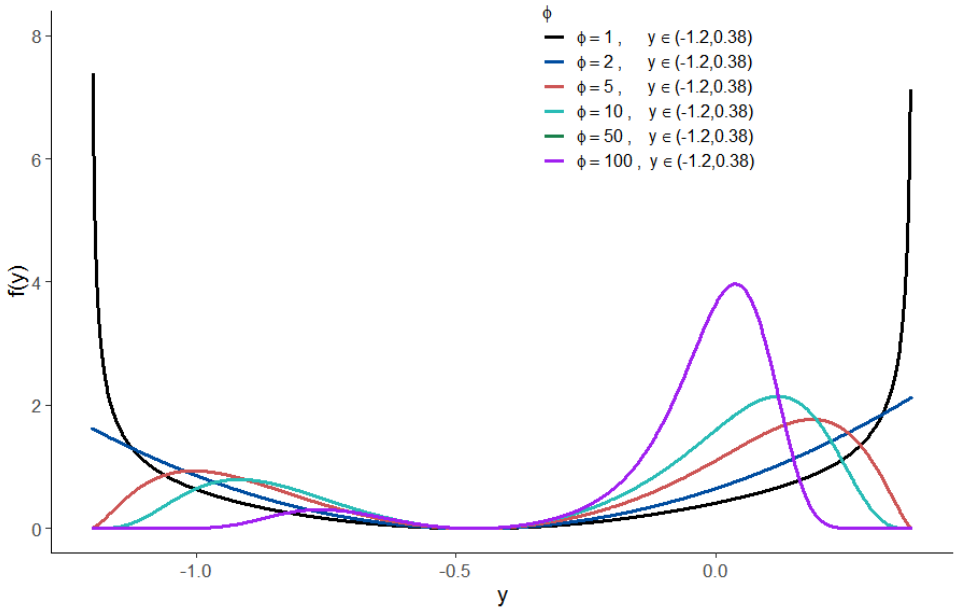
Em síntese, o parâmetro μ_d atua como um parâmetro de localização e corresponde à mediana da distribuição; o parâmetro c parece estar relacionado à assimetria da distribuição e para valores de c muito próximos a 0 ou -1, a distribuição tende a ficar unimodal; o parâmetro μ parece estar próximo da maior moda da distribuição; o parâmetro ϕ parece estar relacionado à precisão da distribuição; o parâmetro δ parece estar relacionado ao distanciamento entre as modas da distribuição, quando há bimodalidade.

Figura 3.3: Curvas da FDP da distribuição BAR segundo variação dos valores de μ, fixados os demais parâmetros ($\phi = 5,\ c = -0,6,\ \mu_d = 0,\ \delta = 2$).



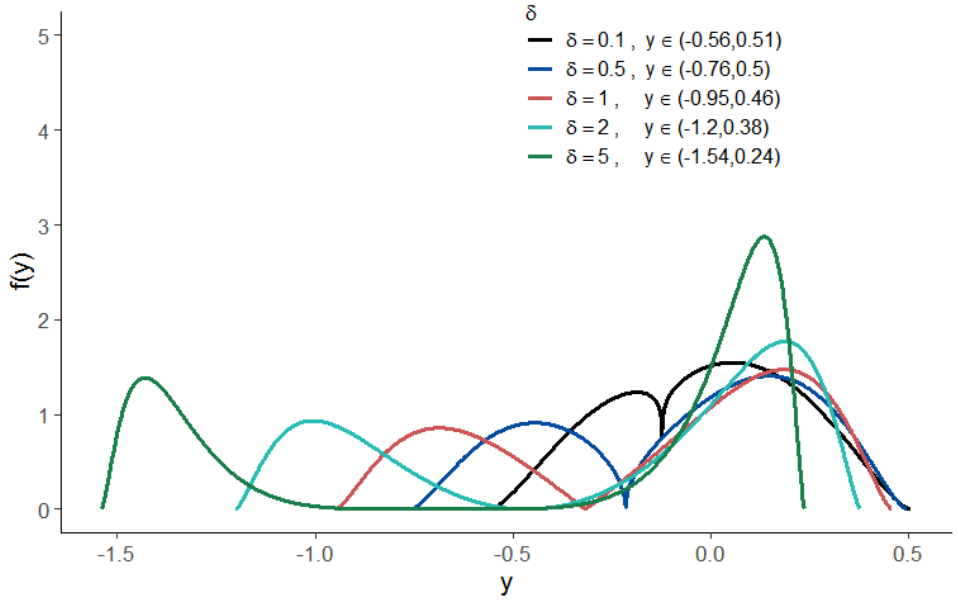
Fonte: Elaboração própria.

Figura 3.4: Curvas da FDP da distribuição BAR segundo variação dos valores de ϕ, fixados os demais parâmetros ($\mu = 0,5,\ c = -0,4,\ \mu_d = 0,\ \delta = 2$).



Fonte: Elaboração própria.

Figura 3.5: Curvas da FDP da distribuição BAR segundo variação dos valores de δ, fixados os demais parâmetros ($\mu = 0,5,\ \phi = 5,\ c = -0,4,\ \mu_d = 0$).



Fonte: Elaboração própria.

Capítulo 4

Modelos de regressão BAR

Quando há interesse em se estudar o comportamento de uma variável resposta, pode-se utilizar os modelos de regressão, uma técnica estatística que permite identificar e quantificar os efeitos de uma ou mais variáveis explicativas na média da variável resposta, bem como prever valores da resposta com base em novos valores das variáveis explicativas (Montgomery, Peck e Vining, 2021). Considerando que os modelos de regressão permitem identificar os sentidos das relações existentes entre as variáveis explicativas e a variável resposta, é de suma importância que essas relações possuam interpretações de fácil entendimento, a fim de se obter uma compreensão mais clara e objetiva acerca do efeito das variáveis explicativas na resposta. Nesse sentido, este capítulo propõe uma nova classe de modelos de regressão utilizando a distribuição beta ampliada reparametrizada como a distribuição condicional postulada para a variável resposta.

4.1 Regressão beta ampliada reparametrizada

Considerando o desenvolvimento da distribuição beta ampliada no capítulo anterior, esta seção trata da definição da estrutura de regressão associada aos parâmetros da distribuição proposta. Importante destacar que a definição da função quantílica da distribuição beta ampliada, em (3.5), permite isolar o parâmetro m de tal forma que a FDA da distribuição BAR fique em

função do quantil de ordem p (τ_p). No contexto de modelos de regressão, essa reparametrização permitirá a interpretação simples dos coeficientes da regressão em termos dos quantis da distribuição BAR. Em particular, consideraremos a modelagem da mediana μ_d da distribuição BAR em função de variáveis explicativas.

A classe de modelos de regressão beta ampliada proposta é definida por

(i) $Y_i \overset{\text{ind}}{\sim} \mathrm{BAR}(\mu_i, \phi_i, c_i, \mu_{di}, \delta),\ i = 1, \ldots, n;$

(ii) $g_{\mu_d}(\mu_{d_i}) = \boldsymbol{x}_i^\top \boldsymbol{\beta} \Leftrightarrow \mu_{d_i} = g_{\mu_d}^{-1}(\boldsymbol{x}_i^\top \boldsymbol{\beta});$

(iii) $g_\phi(\phi_i) = \boldsymbol{z}_i^\top \boldsymbol{\gamma} \Leftrightarrow \phi_i = g_\phi^{-1}(\boldsymbol{z}_i^\top \boldsymbol{\gamma});$

(iv) $g_\mu(\mu_i) = \boldsymbol{v}_i{}^\top \boldsymbol{\eta} \Leftrightarrow \mu_i = g_\mu^{-1}(\boldsymbol{v}_i{}^\top \boldsymbol{\eta});$

(v) $g_c(c_i) = \boldsymbol{w}_i{}^\top \boldsymbol{\lambda} \Leftrightarrow c_i = g_c^{-1}(\boldsymbol{w}_i{}^\top \boldsymbol{\lambda}),$

em que $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{p_1})^\top \in \mathbb{R}^{p_1}$, $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{p_2})^\top \in \mathbb{R}^{p_2}$, $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_{p_3})^\top \in \mathbb{R}^{p_3}$ e $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_{p_4})^\top \in \mathbb{R}^{p_4}$ são vetores de parâmetros de regressão desconhecidos ($p_1 + p_2 + p_3 + p_4 + 1 = p < n$), $\boldsymbol{x}_i{}^\top = (x_{i1}, \ldots, x_{ip_1}) \in \mathbb{R}^{p_1}$, $\boldsymbol{z}_i{}^\top = (z_{i1}, \ldots, z_{ip_2}) \in \mathbb{R}^{p_2}$, $\boldsymbol{v}_i{}^\top = (v_{i1}, \ldots, v_{ip_3}) \in \mathbb{R}^{p_3}$, $\boldsymbol{w}_i{}^\top = (w_{i1}, \ldots, w_{ip_4}) \in \mathbb{R}^{p_4}$ são os vetores de valores fixados para a i-ésima observação nos submodelos para μ_d, ϕ, μ e c, respectivamente, $g_{\mu_d}(\cdot)$, $g_\phi(\cdot)$, $g_\mu(\cdot)$ e $g_c(\cdot)$ são as respectivas funções de ligação para os quatro submodelos, sendo estas ligações monótonas e diferenciáveis.

Por simplicidade, não será atribuída estrutura de regressão ao parâmetro δ e este também não será estimado, decidiu-se por fixá-lo em $\delta = 0{,}5$. Entre as razões para fixar o parâmetro δ, tem-se o fato de que a implementação do modelo utilizará um pacote do software `R` que considera a estimação pelo método de máxima verossimilhança para distribuições de até quatro parâmetros. Também tem-se que δ está no suporte da distribuição, o que pode complicar o procedimento de estimação.

Cabe destacar que, embora a classe de modelos de regressão beta ampliada considere estrutura de regressão para os quatro parâmetros, ao se ajustar um conjunto de dados, não é necessá-

rio atribuir regressão a todos eles. Isso significa que, a depender dos dados, pode-se especificar uma estrutura de regressão apenas para a mediana μ_d desde que o ajuste aos dados se mostre satisfatório. E, pode-se adicionar covariáveis aos demais submodelos, conforme necessário, dependendo da flexibilidade requerida pelo modelo.

Ao atribuir estruturas de regressão aos parâmetros μ_d, ϕ, μ e c estaremos flexibilizando o ajuste do modelo de regressão BAR aos dados. Em particular, a estrutura de regressão atribuída ao parâmetro μ_d permitirá identificar covariáveis que afetam o comportamento da mediana da resposta. Além disso, a estimativa de μ_d poderá ser utilizada para realizar previsões da resposta com base em novos valores fixados das covariáveis. Os demais parâmetros apenas alteram a forma da distribuição e não possuem interpretação prática. Por essa razão, é importante frisar que não se deve confundir a interpretação dos parâmetros μ e ϕ do modelo BAR com aquelas referentes ao modelo beta. Enquanto no modelo BAR μ e ϕ estão associados apenas à forma da distribuição, no modelo beta proposto por Ferrari e Cribari-Neto (2004), μ representa a média e ϕ, a precisão da resposta. Apesar da análise gráfica da Figura 3.4 apresentar indicativos de que parâmetro ϕ esteja relacionado à precisão da distribuição BAR, essa afirmação necessita de comprovação matemática.

Com relação a escolha das funções de ligações no modelo de regressão BAR, estas usualmente são escolhidas de modo que se satisfaça $g_{\mu_d}(\mu_{di}) \in \mathbb{R}$, $g_\phi(\phi_i) \in \mathbb{R}$, $g_\mu(\mu_i) \in \mathbb{R}$, e $g_c(c_i) \in \mathbb{R}$. Estas condições são impostas para garantir que os respectivos preditores lineares em cada submodelo da regressão sejam estimados sem restrições, ou seja, sem limitações sobre os valores que os coeficientes de regressão podem assumir, garantindo que as respectivas estimativas dos parâmetros da distribuição BAR pertençam aos respectivos espaços paramétricos.

Adicionalmente, ressalta-se a possibilidade de escolher funções de ligação aplicadas ao parâmetro modelado que não estejam livres em toda a reta real. Por exemplo, ao modelar uma variável que possua suporte no intervalo $(0, 1)$, pode-se utilizar a função de ligação logit, que restringirá os valores da mediana μ_d a este mesmo intervalo.

Para a modelagem dos parâmetros μ e ϕ da distribuição BAR utilizam-se as funções ligações

enumeradas na Seção 2.3. Em especial, as ligações logit e logarítmica são mais frequentemente utilizadas para μ e ϕ, respectivamente. Tendo em vista que $\mu_d \in \mathbb{R}$, a função de ligação $g_{\mu_d}(\cdot)$ que inicialmente propomos a ser utilizada é a ligação identidade, ou seja, $g_{\mu_d}(\mu_d) = \mu_d$. A interpretação desta função de ligação é direta sendo β_j a variação causada na mediana da resposta ao aumentar uma unidade na variável explicativa x_j, desde que as demais covariáveis estejam fixadas. Destaca-se que utilizando a beta bimodal proposta por Vila et al. (2024) não é possível obter uma interpretação simples como a mencionada.

No contexto de modelos lineares generalizados propostos por Nelder e Wedderburn (1972), a modelagem de parâmetros que assumem valores reais pode ser feita por meio de outras funções de ligações além da função identidade. Dentre as ligações mais utilizadas, destacam-se:

- função recíproca: $g_{\mu_d}(\mu_d) = \frac{1}{\mu_d}$;
- função recíproca ao quadrado: $g_{\mu_d}(\mu_d) = \frac{1}{\mu_d^2}$; ;
- função logit: $g_{\mu_d}(\mu_d) = \log\left(\frac{\mu_d}{1-\mu_d}\right)$.

Cabe destacar que a função de ligação logit para a mediana da distribuição BAR foi elencada por se adequar ao contexto em que a variável resposta assume valores no intervalo $(0, 1)$, visto que a mediana deve ser mapeada nesse mesmo intervalo.

Quanto ao parâmetro c, considerando que estamos interessados em c contido no intervalo $(-1, 0)$, propõe-se a utilização de uma nova função de ligação que mapeia $c \in (-1, 0)$ aos $\mathbb{R}$, denotada por "logit inversa". Essa é a função de ligação atribuída ao parâmetro c que é proposta neste trabalho. A prova de que a função logit inversa é monótona e diferenciável está disponível no Apêndice A. Vale ressaltar que o interesse deste trabalho reside em estimar $c \in (-1, 0)$, contudo, para aplicações da distribuição BAR com foco em $c \in (-1, 1)$, por exemplo, sugere-se verificar a função de ligação "$[-1, 1]$" apresentada por Rigby et al. (2019). Ambas as funções são descritas a seguir:

- função logit inversa: $g_c(c) = \log\left(\frac{-c}{(1+c)}\right)$, $c \in (-1, 0)$;

- função [-1,1]: $g_c(c) = \log\left(\frac{(c-1)}{(1-c)}\right)$, $c \in [-1, 1]$.

4.1.1 Estimação dos parâmetros

Considerando o modelo de regressão BAR proposto, pretende-se estimar os parâmetros que indexam a distribuição $\text{BAR}(\mu, \phi, c, \mu_d, \delta)$, com $\delta = 0{,}5$. Isso significa que o vetor de parâmetros $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \boldsymbol{\gamma}^\top, \boldsymbol{\eta}^\top, \boldsymbol{\lambda}^\top)^\top \in \mathbb{R}^{p_1+p_2+p_3+p_4}$ associado ao modelo de regressão BAR é desconhecido e deve ser estimado com base na amostra. Para tanto, será utilizado o método de máxima verossimilhança para obtenção do estimador de $\boldsymbol{\theta}$, denotado por $\widehat{\boldsymbol{\theta}}$, que maximiza o logaritmo da função de verossimilhança dado por

$$\begin{aligned}
\ell(\boldsymbol{\theta} \mid \boldsymbol{y}) &= n\log(0{,}5+1) - \sum_{i=1}^{n}\log(B(\mu_i\phi_i, (1-\mu_i)\phi_i)) \\
&+ \sum_{i=1}^{n}(\mu_i\phi_i - 1)\log\left[(y_i - m_i)|y_i - m_i|^{0,5} - c_i\right] \\
&+ \sum_{i=1}^{n}\Big((1-\mu_i)\phi_i - 1\Big)\log\left[1 - (y_i - m_i)|y_i - m_i|^{0,5} + c_i\right] \\
&+ 0{,}5\sum_{i=1}^{n}\log(|y_i - m_i|)\}, \qquad (4.1)
\end{aligned}$$

em que $m_i = \mu_{d_i} - \text{sgn}\Big(c_i + F^{-1}(y_i; \mu_i\phi_i, (1-\mu_i)\phi_i)\Big)\Big|c_i + F^{-1}(y_i; \mu_i\phi_i, (1-\mu_i)\phi_i)\Big|^{1/(0,5+1)}$.

Destaca-se que a verossimilhança expressa em (4.1) está bem definida apenas no suporte da distribuição, ou seja, para cada y_i pertencente ao i-ésimo suporte, determinado pelos parâmetros indexados em i.

Ao igualar as derivadas do logaritmo da função verossimilhança (4.1) a zero, obtém-se um sistema complexo de equações não lineares em $\boldsymbol{\theta}$ que não possui solução explícita para $\hat{\boldsymbol{\theta}}$. Por esse motivo, $\hat{\boldsymbol{\theta}}$ foi obtido numericamente por meio da função `gamlss` do pacote `gamlss` do software `R`, uma vez que a implementação do procedimento de estimação dos parâmetros asso-

ciados ao modelo de regressão BAR foi realizada utilizando a estrutura de Modelos Aditivos Generalizados para Localização, Escala e Forma (GAMLSS), proposta por Rigby e Stasinopoulos (2005).

A classe dos GAMLSS é especialmente útil, pois permite que a variável resposta siga qualquer distribuição de até quatro parâmetros e que todos os parâmetros sejam modelados em função de variáveis explicativas, conferindo maior flexibilidade ao modelo (Stasinopoulos et al., 2017). Entre outras vantagens dos GAMLSS, Stasinopoulos et al. (2017) também destaca a possibilidade de inclusão de uma variedade de termos aditivos nos modelos para os parâmetros, e a extensão de modelos estatísticos básicos, permitindo modelagem de sobredispersão, excesso de zeros, assimetria e curtose nos dados.

Segundo Rigby e Stasinopoulos (2005), existem dois algoritmos para estimação dos parâmetros sob os GAMLSS. O algoritmo CG, que é uma generalização do algoritmo de Cole e Green (1992); e o RS, uma generalização do algoritmo proposto por Rigby e Stasinopoulos (1996) no ajuste de modelos MADAM (Modelos Aditivos para média e dispersão). Em geral, o método RS é mais estável e mais rápido que o CG, contudo quando os parâmetros da distribuição são correlacionados, o algoritmo RS pode ser mais lento e convergir antes de atingir o logaritmo da função de verossimilhança máxima global (Stasinopoulos et al., 2017).

Complementarmente, outra propriedade interessante dos GAMLSS consiste na possibilidade de especificar os valores iniciais dos algoritmos diretamente para os parâmetros da distribuição condicional da resposta, ao invés de para os parâmetros das estruturas de regressão, o que torna a implementação mais simples (Stasinopoulos et al., 2017). A esse respeito, para o modelo aqui proposto, propõe-se a mediana da variável resposta como valor inicial para μ_d; a estimativa do parâmetro ϕ resultante do ajuste da regressão beta para ϕ; e $0{,}5$ para μ, e $-0{,}5$ para c.

4.1.2 Avaliação e seleção de modelos, teste de hipóteses e intervalos de confiança

Para avaliar a qualidade do ajuste dos modelos de regressão BAR, os resíduos quantílicos, introduzidos por Dunn e Smyth (1996), definidos por

$$r_{q,i} = \Phi^{-1}\left\{F(y_i; \hat{\mu}_i, \hat{\phi}_i, \hat{\mu_{di}}, \hat{c}_i, 0{,}5)\right\},$$

em que $\Phi(\cdot)$ é a FDA da distribuição normal padrão e $F(y; \mu, \phi, \mu_d, c, 0{,}5)$ é a FDA da distribuição BAR. Se o modelo de regressão BAR estiver bem ajustado, $r_{q,i}$ para $i = 1, \ldots, n$ seguirão distribuição normal padrão e serão independentes para n grande.

Para avaliar os resíduos quantílicos, pode-se utilizar diversos gráficos, como, por exemplo, o *worm plot* proposto por Buuren e Fredriks (2001). Este consiste em um gráfico nos moldes de um QQplot usual, entretanto, sem tendência. Nele, o eixo das abscissas apresenta os quantis esperados considerando o modelo com ajuste adequado, e o eixo das ordenadas exibe os desvios (diferença entre os valores dos resíduos quantílicos e dos quantis esperados da normal padrão). Esse gráfico também apresenta as bandas de confiança baseadas na normalidade assintótica dos resíduos quantílicos. Caso o modelo postulado apresente um ajuste adequado, espera-se que os pontos estejam próximos à linha centrada em zero e que no máximo 5% deles estejam fora da região limitada pelas bandas de confiança. Por outro lado, a presença de padrões sistemáticos sugere inadequação do ajuste, podendo indicar deficiências nos resíduos e na distribuição ajustada pelo modelo. Nesse sentido, a Tabela 4.1, apresentada por Stasinopoulos et al. (2017), associa determinados formatos dos pontos a essas falhas. A partir disso, pode-se ter indícios acerca do comportamento da variável resposta e de como aprimorar o ajuste aos dados.

Complementarmente, a escolha entre modelos considerados adequados pode ser apoiada pelo Critério de Akaike (AIC), proposto por Akaike (1974) e definido como

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

Tabela 4.1: Formas típicas do *worm plot* dos resíduos e suas implicações sobre o modelo ajustado.

Formato do *worm plot* (ou da curva ajustada)	**Resíduos**	**Interpretação na Distribuição ajustada**
Concentração acima da origem	Média muito alta	Parâmetro de localização subestimado
Concentração abaixo da origem	Média muito baixa	Parâmetro de localização ajustada muito alta
Inclinação positiva	Variância muito alta	Parâmetro de escala muito baixo
Inclinação negativa	Variância muito baixa	Parâmetro de escala muito alta
Formato em U	Assimetria à direita	Excesso de assimetria à esquerda
Formato em U invertido	Assimetria à esquerda	Excesso de assimetria à direita
Formato em S com parte esquerda voltada para baixo	Leptocúrticos	Caudas muito leves
Formato em S com parte esquerda voltada para cima	Platicúrticos	Caudas muito pesadas

Fonte: adaptado de Stasinopoulos, Rigby e Heller (2017, p. 428).

em que $\ell(\hat{\boldsymbol{\theta}})$ é o logaritmo da função de verossimilhança de $\boldsymbol{\theta}$ avaliado em $\hat{\boldsymbol{\theta}}$, e k é a quantidade de parâmetros no modelo ajustado. Por definição, o modelo ajustado que apresenta menor AIC é preferível. Burnham e Anderson (2002) recomendam utilizar o AIC em situações nas quais o tamanho da amostra é, ao menos, 40 vezes superior à quantidade de parâmetros. Segundo Davison (2003), o AIC tende a superparametrizar os modelos ao selecionar especificações mais complexas que o necessário.

Com a finalidade de investigar a significância estatística dos coeficientes da regressão, ou seja, sua relevância para a composição do preditor linear da regressão, pode-se utilizar testes de hipóteses. Especificamente, para o método de máxima verossimilhança, utiliza-se o teste de Wald. Conforme abordado em Montgomery, Peck e Vining (2021), as hipóteses a serem testadas para um coeficiente β_j são $H_0 : \beta_j = 0$ contra $H_1 : \beta_j \neq 0$. Se H_0 não for rejeitada, a um nível de significância $\alpha 100\%$, $0 < \alpha < 1$, ou seja, se não há evidência para rejeitar a hipótese nula de que $\beta_j = 0$, não há indícios de que a covariável x_j tenha efeito na variável resposta. A estatística do teste é dada por:

$$z = \frac{\hat{\beta}_j}{ep(\hat{\beta}_j)}$$

em que $ep(\hat{\beta}_j)$ é o erro-padrão de β_j.

Para obtenção dos intervalos de confiança aproximados dos parâmetros, considere uma distribuição com vetor de k parâmetros $\boldsymbol{\theta}$ e estimadores de máxima verossimilhança (EMVs) $\hat{\boldsymbol{\theta}}$. Segundo Rigby et al. (2019), considerando a normalidade assintótica dos EMVs, a distribuição assintótica de $\hat{\theta_k}$ quando $n \to \infty$ é

$$\hat{\theta_k} \overset{a}{\sim} N(\theta_k, [ep(\hat{\theta_k})]^2),$$

em que o erro padrão, $ep(\hat{\theta_k})$, é a raiz quadrada do k-ésimo elemento de $[\boldsymbol{i}(\boldsymbol{\theta})]^{-1}$, sendo $\boldsymbol{i}(\boldsymbol{\theta})$ a matriz de informação de Fisher observada. Então, o intervalo de confiança de $100(1-\alpha)\%$ para um determinado θ_k é, aproximadamente:

$$\left(\hat{\theta_k} \pm z_{\alpha/2}\widehat{ep}(\hat{\theta_k})\right),$$

em que $z_{\alpha/2}$ o quantil da distribuição normal padrão associado à probabilidade $\alpha/2$ da cauda superior e $\widehat{ep}(\hat{\theta_k})$, é a raiz quadrada do k-ésimo elemento da diagonal da matriz $[\boldsymbol{i}(\boldsymbol{\theta})]^{-1}$, sendo $\boldsymbol{i}(\boldsymbol{\theta})$ a matriz de informação de Fisher avaliada nas estimativas de máxima verossimilhança.

Capítulo 5

Aplicação a dados reais

5.1 Descrição dos dados

Em tempos contemporâneos, o debate sobre questões ambientais e sustentabilidade tem tomado destaque, principalmente por estarem diretamente relacionadas à subsistência humana. Nesse contexto, entre as mobilizações relacionadas ao tema, pode-se citar a realização da Conferência das Nações Unidas sobre as Mudanças Climáticas - COP30, que objetiva o debate global sobre ações de combate à crise climática[1]; o Plano de Ação para a Saúde de Belém, proposta brasileira para ajudar os países a adaptarem seus sistemas de saúde aos efeitos das mudanças climáticas[2]; e o desenvolvimento de um protocolo de calor pela Prefeitura do Rio de Janeiro, que utiliza o Índice de Calor, obtido pela combinação entre temperatura média da cidade e umidade relativa do ar[3].

Nesse sentido, estudos relacionados a condições climáticas são de extrema relevância para compreender como se dão e como os fenômenos se relacionam. Sendo assim, para exemplificar a aplicabilidade do modelo BAR, proposto neste trabalho, considerou-se um conjunto de dados climáticos extremos da estação automática A002, situada na cidade de Goiânia, Goiás, ocorridos

[1] https://cop30.br/pt-br/sobre-a-cop30. Acesso em 21/11/2025.

[2] https://www.gov.br/saude/pt-br/assuntos/cop30/publicacoes/plano-de-acao-em-saude-de-belem-portugues.pdf. Acesso em 21/11/2025.

[3] https://saude.prefeitura.rio/protocolo-de-calor/.Acesso em 26/10/2025.

entre 01/01/2011 e 31/12/2022, tais dados também foram analisados no trabalho de Lisboa (2024). Os dados são disponibilizados pelo Instituto Nacional de Meteorologia e possuem, ao todo, 73 observações para as seguintes variáveis:

- **Umidade Relativa Mínima do Ar (URA):** Razão entre a quantidade de vapor de água presente numa porção da atmosfera e a quantidade máxima de vapor de água que a atmosfera pode suportar (saturação) a uma mesma temperatura e pressão[4]. Expressa em porcentagem;
- **Temperatura Mínima do Orvalho (TO):** Temperatura à qual o ar deve ser resfriado para atingir a saturação, formando orvalho [5]. Medida em graus Celsius (Cº).;
- **Pressão Atmosférica Média (PAM):** É a pressão atmosférica média da última hora anterior à mensagem de dados . A pressão atmosférica é definida como a força exercida pelo peso da atmosfera sobre qualquer superfície terrestre [6]. Medida em milibares (mb);
- **Rajada Máxima de vento (RV):** É a velocidade máxima do vento ocorrida na última hora anterior à mensagem de dados[7]. Expressa em metros por segundo (m/s);
- **Estação (E):** Variável indicadora de estação chuvosa (correspondendo aos meses de outubro a abril), assumindo valor 1 nesse período e 0 durante os demais meses.

A variável resposta de interesse deste estudo será a "Umidade relativa mínima do ar", uma vez que possui características que podem ser descritas pela distribuição BAR: é contínua; e, seu domínio pertence ao intervalo $(0, 1)$.

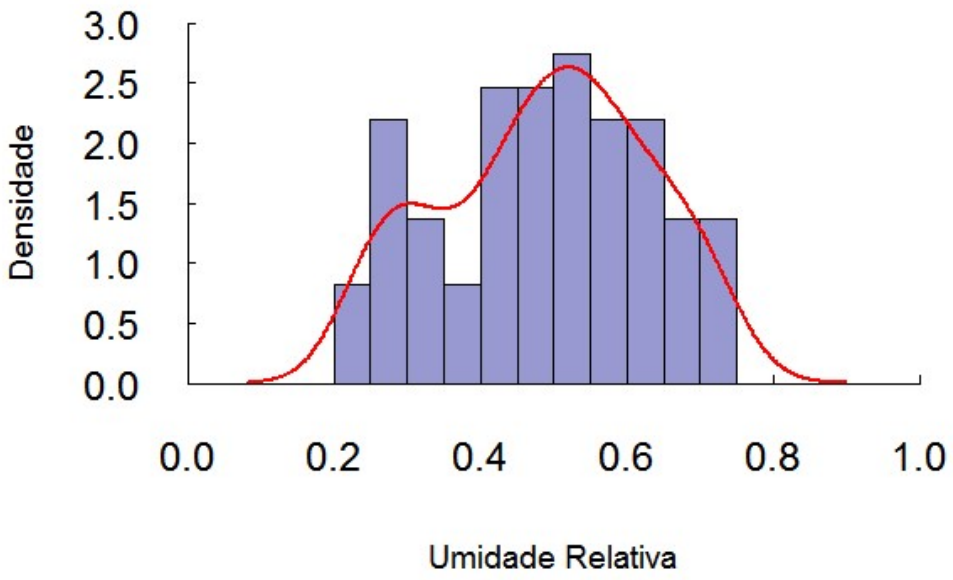
Dessa forma, pretende-se verificar a capacidade das distribuições beta e beta ampliada de acomodar a variável resposta, bem como ajustar um modelo de regressão, considerando a hipótese de que Umidade $\sim \mathrm{BAR}(\mu, \phi, \mathrm{c}, \mu_{\mathrm{d}}, \delta = 0{,}5)$.

[4] https://climaesaude.icict.fiocruz.br/indicador/umidade-relativa-do-ar-pontual. Acesso em 24/11/2025.
[5] https://portal.inmet.gov.br/glossario/glossario. Acesso em 24/11/2025.
[6] https://portal.inmet.gov.br/glossario/glossario. Acesso em 24/11/2025.
[7] https://portal.inmet.gov.br/glossario/glossario. Acesso em 24/11/2025.

5.2 Análise Descritiva

Esta seção trata da análise descritiva das variáveis disponíveis no conjunto de dados. Com essa etapa, espera-se detalhar o comportamento da variável resposta, bem como identificar os possíveis relacionamentos com as demais variáveis. A Figura 5.1 apresenta o histograma da umidade relativa do ar juntamente com a curva de densidade estimada (não paramétrica) pelo comando `density` do software `R`. A Tabela 5.1 apresenta as principais estatísticas descritivas da variável umidade relativa.

Figura 5.1: Histograma e curva de densidade estimada da Umidade relativa do ar.



Fonte: Elaboração própria.

O histograma de URA revela um comportamento bimodal, o qual é acompanhado pela densidade estimada. Também observa-se uma maior concentração de valores no intervalo $[0,4; 0,6]$, o que é positivo, pois essa faixa de umidade relativa do ar apresenta menor risco de proliferação de microrganismos e de infecções respiratórias (Jones et al., 2022). Conforme a Tabela 5.1, a URA varia de 0,24 a 0,74, aproximadamente, e apresenta média e mediana amostrais próximas, sendo a mediana superior. Além disso, a distribuição é platicúrtica, uma vez

Tabela 5.1: Estatísticas descritivas para Umidade relativa do ar.

Estatística	**Valor**
Mínimo	0,240
1º Quartil	0,385
Mediana	0,498
Média	0,489
3º Quartil	0,587
Máximo	0,739
Assimetria	-0,178
Curtose	2,063
Número de observações	73

Fonte: Elaboração própria.

que a curtose é menor que 3 (padrão mesocúrtico); e levemente assimétrica à esquerda (coeficiente de assimetria negativo e próximo a zero). Essas medidas refletem na assimetria dos picos (modas) da distribuição, sendo o da esquerda menor.

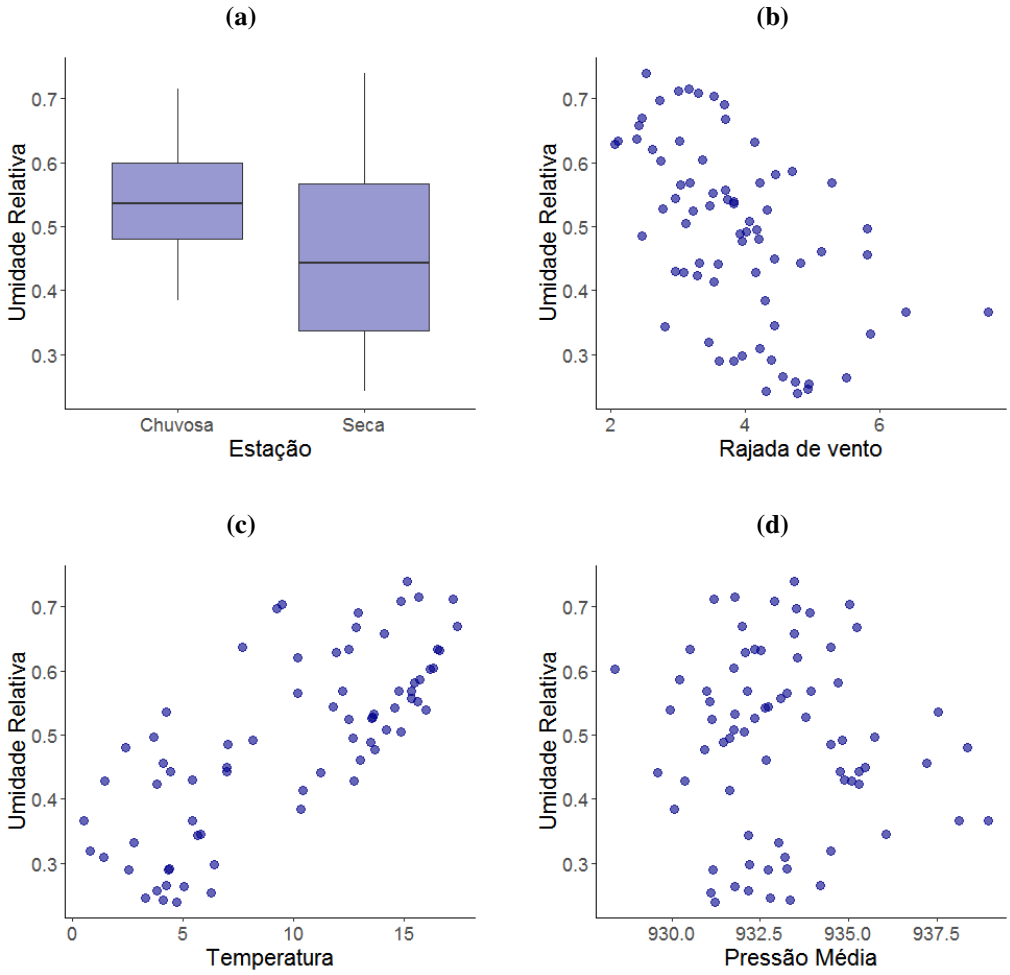
Com o intuito de identificar as candidatas a variáveis explicativas que possam estar correlacionadas à variável resposta (URA) e para identificar padrões de relacionamento entre elas, realizou-se a análise gráfica apresentada na Figura 5.2, que reúne o boxplot de URA segundo estacão, bem como os gráficos de dispersão de URA segundo RV, TO, PAM. Nesse sentido, a Figura 5.2a mostra que a estação chuvosa apresentou maior mediana amostral de URA, enquanto a estação seca apresentou maior variabilidade, uma vez apresenta uma caixa maior. Percebe-se, também, que URA possui uma relação, aparentemente linear, positiva com TO (5.2c) e negativa com RV (5.2b), no entanto, não apresenta um padrão claro de associação com PAM (5.2d).

Uma boa prática ao se trabalhar com modelos de regressão linear é investigar a existência de multicolinearidade. Isso porque tal fenômeno implica em uma dependência quase linear entre os regressores, que pode acarretar em problemas nas estimativas do modelo proposto (Montgomery, Peck e Vining, 2021). Nesse sentido, para investigar se há indícios de correlação linear entre as candidatas a explicativas, foram elaborados os gráficos de dispersão das variáveis candidatas a explicativas umas contra as outras. Esses são apresentados na Figura 5.3.

Analisando a Figura 5.3, verifica-se que não há indício de relação linear entre RV e TO

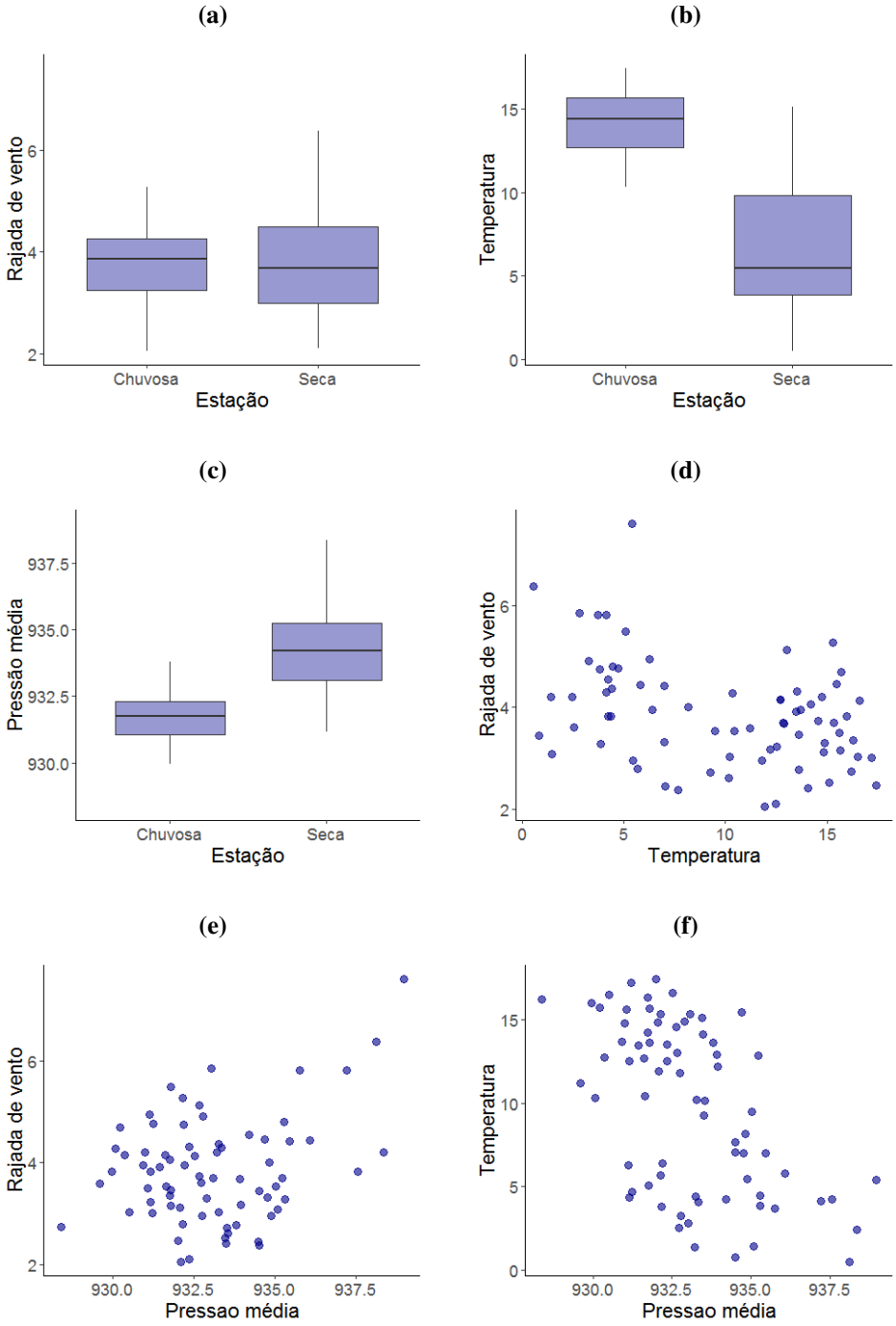
(5.3d), tampouco entre RV e PAM (apesar da Figura 5.3e apresentar alguns valores discrepantes, a maior concentração de pontos não possui padrão). Contudo, observa-se o oposto para TO e PAM, o que deve ser levado em consideração durante a avaliação do ajuste e na seleção de variáveis do modelo de regressão. Além disso, observa-se que a distribuição de RV apresenta comportamento similar para diferentes valores de E. No entanto, E parece estar associada à TO e PAM, sendo a relação com TO, mais expressiva, tendo em vista as diferenças entre as medianas e formas aparentes da distribuição de TO para estação chuvosa e seca.

Figura 5.2: Umidade relativa do ar versus variáveis explicativas.



Fonte: Elaboração própria.

Figura 5.3: Relações entre candidatas a variáveis explicativas.



Fonte: Elaboração própria.

5.3 Ajuste marginal da distribuição

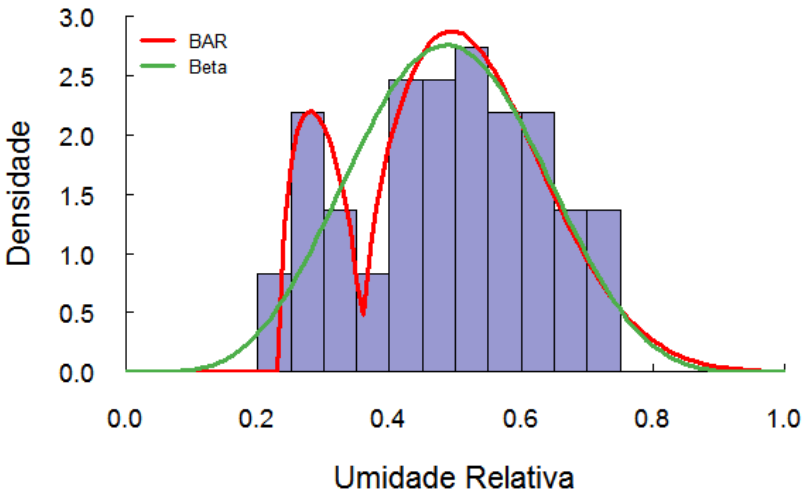
A Umidade relativa do ar é uma variável contínua, cujo domínio pertence ao intervalo $(0, 1)$. Em especial, a análise descritiva revelou que a função densidade da URA medida na cidade de Goiânia entre 2011 e 2022 (objeto de estudo desta aplicação) apresenta bimodalidade. Portanto, ajustar distribuições que admitem apenas comportamentos unimodais pode não ser apropriado. Por essa razão, espera-se que a distribuição BAR seja mais adequada para descrever o comportamento da variável resposta estudada do que a distribuição beta.

A fim de avaliar a capacidade de ajuste da distribuições beta e BAR na modelagem de URA, ajustou-se uma regressão BAR considerando apenas o intercepto em todos os submodelos, isto é, sem inclusão de variáveis explicativas (covariáveis). E, para fins de comparação, foi realizado o mesmo procedimento com a regressão beta, também considerando o ajuste independente e identicamente distribuído (IID). Cabe ressaltar que todos os ajustes subsequentes foram realizados utilizando a função `gamlss` do software R. Em particular, a família BE considera a reparametrização da distribuição beta proposta por Rigby et al. (2019), ou seja, os parâmetros estimados são μ e σ, que correspondem à média e à dispersão da distribuição beta, respectivamente, com ambos pertencendo ao intervalo (0,1). Para o ajuste do modelo BAR, fixou-se o parâmetro $\delta = 0{,}5$ e utilizou-se a família desenvolvida nesta dissertação. Os resultados correspondentes são apresentados na Figura 5.4 e nas Tabelas 5.2 e 5.3. A Figura 5.4 apresenta o histograma de URA, juntamente com as curvas de densidade estimadas pelo modelos beta e BAR especificados. A Tabela 5.2 apresenta os valores do Critério de Akaike, enquanto a Tabela 5.3 apresenta as estimativas e os erros padrões dos modelos ajustados de URA.

A Figura 5.4 indica que a distribuição BAR se ajusta melhor aos dados quando comparada com a distribuição beta, pois captura a bimodalidade empírica, e a curva estimada em vermelho apresenta maior aderência ao histograma do que a curva em verde, que é essencialmente unimodal. Esse resultado concorda com os AICs exibidos na Tabela 5.2, uma vez que o ajuste BAR possui o menor AIC. Sendo, portanto, o escolhido para descrever o comportamento marginal de

URA. Ao analisar a Tabela 5.3, percebe-se que os valores das estimativas dos parâmetros das distribuições são próximos em ambos os ajustes. É natural que não sejam exatamente iguais, tendo em vista que se tratam de estimativas de modelos diferentes, em que a regressão Beta estima parâmetros relacionados à média e dispersão; enquanto a regressão BAR estima a mediana e outros três parâmetros de forma. No entanto, as estimativas próximas são plausíveis, visto que média e mediana são parâmetros de tendência central, cujas respectivas estatísticas se mostraram muito próximas na amostra (Tabela 5.1).

Figura 5.4: Histograma de URA juntamente com as densidades ajustadas via distribuição beta e BAR.



Fonte: Elaboração própria.

Tabela 5.2: Valores de AIC para os ajustes IID das distribuições Beta e BAR.

Distribuição	AIC
Beta	-80,40
BAR	-87,76

Fonte: Elaboração própria.

Tabela 5.3: Estimativas e erros-padrão dos ajustes IID das distribuições Beta e BAR.

Beta	
μ	σ
0,489	0,273
(0,016)	(0,020)

BAR			
μ_d	ϕ	μ	c
0,492	14,723	0,112	-0,045
(0,018)	(4,095)	(0,013)	(0,010)

Fonte: Elaboração própria.

5.4 Ajuste via modelos de regressão

Nesta seção, propõe-se estender a análise iniciada na Seção 2.2, modelando a URA em função das covariáveis disponíveis. Para tanto, serão avaliados os desempenhos dos modelos de regressão beta no ajuste da média e da regressão BAR no ajuste da mediana da variável resposta. O modelo beta utiliza a parametrização de Rigby et al. (2019), considerando a função de ligação logit para a média μ e para a dispersão σ. Enquanto o modelo BAR considera as ligações logit para a mediana μ_d e para μ; logarítmica para ϕ; e, logit inversa para c. Tendo em vista que a variável resposta assume valores no intervalo $(0, 1)$, a ligação logit torna-se uma escolha natural para modelar a mediana da resposta.

A adequação dos modelos será avaliada através da análise dos *worm plots* dos resíduos quantílicos. O *worm plot* é uma ferramenta para avaliar se os resíduos seguem a distribuição proposta que, quando bem ajustados, apresentam os pontos próximos da linha horizontal centrada em zero (Buuren e Fredriks, 2001).

A seleção de variáveis será feita pelo método passo a frente (*forward*). O método *forward* consiste em assumir um modelo inicial e encontrar um modelo com bom ajuste incluindo as covariáveis uma por vez. A partir da análise descritiva, identificaram-se possíveis associações

entre URA e as regressoras candidatas disponíveis no banco de dados, especialmente E, RV e TO. Tendo em vista esse fato, bem como as possíveis associações entre TO e E; TO e PAM ; e entre PAM e E, consideraram-se quatro especificações de modelos. Todas as estruturas de regressão foram ajustadas considerando $Y_i \overset{ind.}{\sim} \text{Beta}(\mu_i, \sigma)$ e $Y_i \overset{ind.}{\sim} \text{BAR}(\mu, \phi, \mu_{di}, c, 0{,}5)$. As especificações são apresentadas a seguir:

i) Especificação 1: $\eta_i = \beta_1 + \beta_2 \text{E}_i$;

ii) Especificação 2: $\eta_i = \beta_1 + \beta_2 \text{E}_i + \beta_3 \text{RV}_i$;

iii) Especificação 3: $\eta_i = \beta_1 + \beta_2 \text{E}_i + \beta_3 \text{RV}_i + \beta_4 \text{TO}_i$;

iv) Especificação 4: $\eta_i = \beta_1 + \beta_2 \text{E}_i + \beta_3 \text{RV}_i + \beta_4 \text{TO}_i + \beta_5 \text{PAM}_i$,

em que, para ambos os modelos de regressão beta e BAR, $i = 1, \ldots, 73$, E_i, RV_i, TO_i e PAM_i são os i-ésimos valores da estação, da rajada máxima de vento, da temperatura do orvalho e da pressão atmosférica média, respectivamente. No modelo de regressão beta, tem-se que $\eta_i = \text{logit}(\mu_i)$, em que μ_i corresponde à i-ésima média de URA. Enquanto no modelo de regressão BAR, $\eta_i = \text{logit}(\mu_{di})$, sendo μ_{d_i} a i-ésima mediana de URA.

A Tabela 5.4 apresenta os resultados dos modelos especificados no que se refere às estimativas dos parâmetros, erro-padrão, estatística z e p-valor. Nota-se que os erro-padrões das estimativas dos coeficientes da regressão na mediana de URA (modelo BAR) são menores que os erro-padrões dos coeficientes da regressão na média de URA (modelo beta), exceto sob a Especificação 1. Também é interessante destacar que o modelo de regressão BAR parece consistente, uma vez que os sinais das estimativas dos coeficientes de regressão associados às covariáveis de cada especificação concordam com os sinais dessas mesmas estimativas sob o modelo beta. Isso quer dizer que as covariáveis afetam a média (modelo beta) e a mediana (modelo BAR) da distribuição da URA no mesmo sentido, o que faz sentido, visto que tanto a média como mediana são medidas de tendência central.

Tabela 5.4: Estimativas, erros-padrão, estatística z e p-valores dos modelos de regressão Beta e BAR ajustados.

Especificação 1									
	Beta					BAR			
	Estimativa	Erro padrão	Estatística z	p-valor		Estimativa	Erro padrão	Estatística z	p-valor
par,					*par,*				
β_1	-0,156	0,086	-1,823	0,073	β_1	-0,018	0,009	-1,930	0,058
β_2	0,237	0,125	1,893	0,063	β_2	0,010	0,011	0,915	0,363
σ	0,267	0,025	10,892	$< 0,001$	ϕ	66,498	11,531	5,767	$< 0,001$
-	-	-	-	-	μ	0,575	0,002	313,972	$< 0,001$
-	-	-	-	-	c	-0,575	0,002	-369,625	$< 0,001$

Especificação 2									
	Beta					BAR			
	Estimativa	Erro padrão	Estatística z	p-valor		Estimativa	Erro padrão	Estatística z	p-valor
par,					*par,*				
β_1	0,957	0,228	4,197	$< 0,001$	β_1	1,075	0,175	6,157	$< 0,001$
β_2	0,213	0,108	1,963	0,054	β_2	0,184	0,044	4,163	$< 0,001$
β_3	-0,287	0,057	-5,074	$< 0,001$	β_3	-0,317	0,021	-15,353	$< 0,001$
σ	0,229	0,017	13,087	$< 0,001$	ϕ	111,156	21,552	5,158	$< 0,001$
-	-	-	-	-	μ	0,610	0,083	7,348	$< 0,001$
-	-	-	-	-	c	-0,610	0,085	-7,138	$< 0,001$

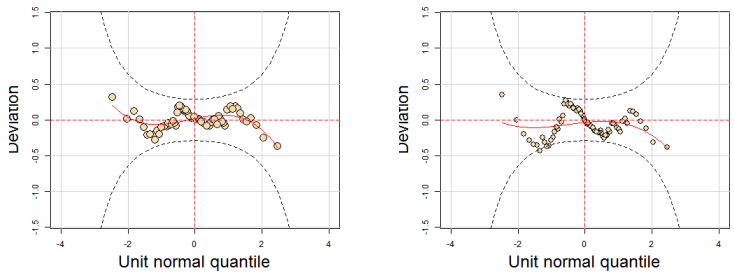
Especificação 3									
	Beta					BAR			
	Estimativa	Erro padrão	Estatística z	p-valor		Estimativa	Erro padrão	Estatística z	p-valor
par,					*par,*				
β_1	-0,546	0,209	-2,614	0,011	β_1	-0,348	0,111	-3,143	0,003
β_2	-0,522	0,104	-5,013	$< 0,001$	β_2	-0,299	0,049	-6,083	$< 0,001$
β_3	-0,095	0,04	-2,342	0,022	β_3	-0,016	0,013	-1,279	0,205
β_4	0,113	0,011	9,948	$< 0,001$	β_4	0,054	0,005	10,185	$< 0,001$
σ	0,149	0,012	12,489	$< 0,001$	ϕ	163,132	44,244	3,687	$< 0,001$
-	-	-	-	-	μ	0,603	0,194	3,107	0,002
-	-	-	-	-	c	-0,600	0,195	-3,078	0,002

Especificação 4									
	Beta					BAR			
	Estimativa	Erro padrão	Estatística z	p-valor		Estimativa	Erro padrão	Estatística z	p-valor
par,					*par,*				
β_1	-117,043	0,235	-498,32	$< 0,001$	β_1	-85,168	0,125	-681,243	$< 0,001$
β_2	-0,242	0,082	-2,941	0,004	β_2	-0,063	0,068	-0,916	0,363
β_3	-0,152	0,031	-4,872	$< 0,001$	β_3	-0,115	0,008	-14,106	$< 0,001$
β_4	0,119	0,009	13,253	$< 0,001$	β_4	0,065	0,001	63,054	$< 0,001$
β_5	0,125	$< 0,001$	586,118	$< 0,001$	β_5	0,091	$< 0,001$	342,445	$< 0,001$
σ	0,117	0,010	12,208	$< 0,001$	ϕ	334,990	155,463	2,155	0,031
-	-	-	-	-	μ	0,606	0,391	1,548	0,122
-	-	-	-	-	c	-0,600	0,390	-1,536	0,125

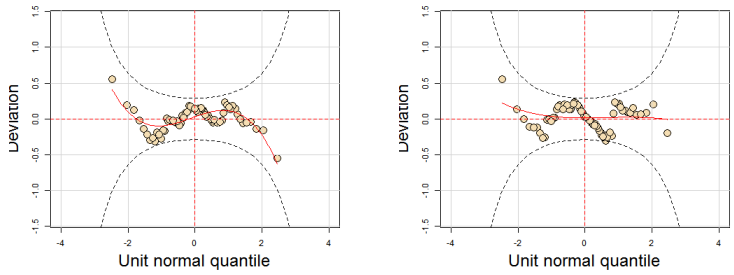
Fonte: Elaboração própria.

Figura 5.5: *Worm plots* dos resíduos quantílicos dos modelos de regressão Beta e BAR ajustados.

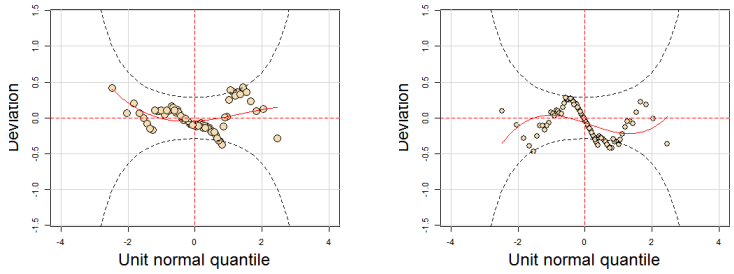
(a) Beta $\sim E_i$ **(b)** BAR $\sim E_i$



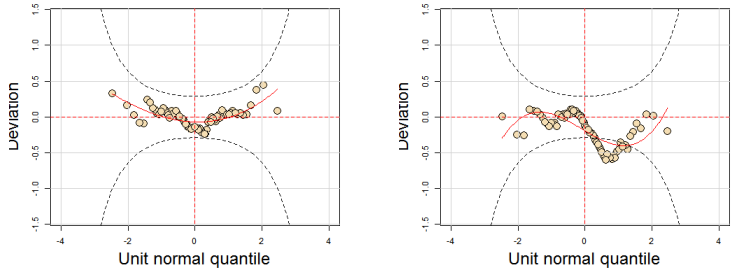
(c) Beta $\sim E_i + RV_i$ **(d)** BAR $\sim E_i + RV_i$



(e) Beta $\sim E_i + RV_i + TO_i$ **(f)** BAR $\sim E_i + RV_i + TO_i$



(g) Beta $\sim E_i + RV_i + TO_i + PAM_i$ **(h)** BAR $\sim E_i + RV_i + TO_i + PAM_i$



Fonte: Elaboração própria.

Ademais, considerando a Tabela 5.4 e o nível de significância de 5%, observa-se que nenhum coeficiente de regressão foi significativo na Especificação 1, tanto no modelo beta quanto no BAR. Na Especificação 2, todos coeficientes do modelo BAR foram significativos, diferentemente do modelo beta. Para a Especificação 3, todos coeficientes foram significativos no modelo beta, enquanto no modelo BAR apenas o coeficiente da variável RV não foi, possivelmente devido à relação entre as variáveis E e TO. Na Especificação 4, todos coeficientes foram significativos no modelo beta; no BAR, somente o intercepto e os coeficientes de RV, TO e PAM. Contudo, cabe lembrar que a relação entre TO, PAM e E pode gerar multicolinearidade, comprometendo a confiabilidade dos coeficientes de ambos os modelos para as especificações 3 e 4.

Os *worm plots* dos resíduos quantílicos resultantes dos modelos de regressão beta e BAR considerando as Especificações 1 a 4 estão dispostos na Figura 5.5. Observa-se que os ajustes considerando as Especificações 1 e 2, baseados na distribuição beta, apresentam resíduos formando curvas vermelhas em formato de S mais acentuado, com a cauda esquerda voltada para cima. Conforme Buuren e Fredriks (2001), esse comportamento indica que há um problema na estimativa da curtose dos resíduos, de modo que a distribuição se torna platicúrtica e a distribuição ajustada possui caudas pesadas (veja a Tabela 4.1). Por outro lado, as curvas atreladas aos modelos 1 e 2, considerando a distribuição BAR, são atenuadas, salientando-se a pertencente ao modelo 2, que tangencia a linha centrada em zero. Os *worm plots* referentes às Especificações 3 e 4 revelam piores ajustes, o que pode estar relacionado à multicolinearidade decorrente da associação entre TO, PAM e E. Verifica-se que a Especificação 2 para o modelo BAR apresentou o melhor ajuste entre todos os 8 modelos de regressão ajustados. Sendo assim, elege-se o modelo 2, por apresentar melhor ajuste.

Considerando o resultado da regressão BAR para a Especificação 2 (Tabela 5.4), ao nível de significância de 5%, observa-se que os coeficientes das variáveis RV e E são significativos para explicar a umidade relativa do ar, considerando a hipótese nula de que os coeficientes da regressão associados sejam iguais a zero ($\beta_j = 0$). Os modelos de regressão BAR foram ajusta-

dos considerando a função de ligação logit para a mediana de URA. Por se tratar de uma função monótona crescente, o aumento (ou diminuição) do logito da mediana da variável resposta implica em aumento (ou diminuição) da mediana da resposta. Dado o modelo final escolhido sob regressão BAR, constata-se que, na estação A002 da cidade de Goiânia, a estação chuvosa possui URA mediana superior à da estação seca (dado o coeficiente $\hat{\beta}_2$ positivo), mantida RV constante. Paralelamente, o aumento da rajada máxima de vento diminui a mediana de URA (coeficiente $\hat{\beta}_3$ negativo), para uma estação fixada. Adicionalmente, a função de ligação logit permite interpretar os coeficientes de regressão em termos da razão de chances mediana da variável resposta, definida por $\mu_d/(1-\mu_d)$. Nesse contexto, entende-se a chance da URA mediana (μ_d) como sendo a razão entre a URA mediana e a proporção que falta para se chegar à saturação do ar (quantidade máxima de vapor que a atmosfera pode suportar). Sem perda de sentido, pode-se traduzir como a chance de umidade relativa mediana aproximar-se da saturação.

Assim, verifica-se que, mantida constante a rajada máxima de vento, durante a estação chuvosa a chance da umidade relativa mínima mediana se aproximar da saturação é 20,2% maior que a da estação seca. Enquanto, mantida a estação constante, a cada acréscimo de 1 m/s na rajada máxima do vento, a chance da umidade relativa mínima mediana se aproximar da saturação é multiplicada por 0,728 ($\mathrm{e}^{-0,317}$), ou seja, diminui em 27,2%.

5.4.1 Ajuste via regressão - modelagem simultânea dos parâmetros

Esta seção estende a aplicação da seção anterior para ilustrar o funcionamento do modelo de regressão BAR com inclusão de regressão em todos os parâmetros da distribuição. É importante destacar que a implementação do modelo de regressão BAR por meio do GAMLSS foi essencial para possibilitar a extensão do ajuste aos demais parâmetros da distribuição, além da mediana da variável resposta, uma vez que o GAMLSS permite a modelagem simultânea de até quatro parâmetros da distribuição. Tendo em vista que o modelo escolhido para explicar a mediana da URA foi aquele em função de E e RV, ajustou-se a Especificação 2 para a mediana, e adicionou-se a covariável RV nos submodelos de regressão para os parâmetros ϕ, μ e c. Os resultados

do ajuste são apresentados na Tabela 5.5. A Figura 5.6 apresenta o *worm plot* dos resíduos quantílicos e, a despeito da maior complexidade do modelo, o ajuste mostrou-se satisfatório.

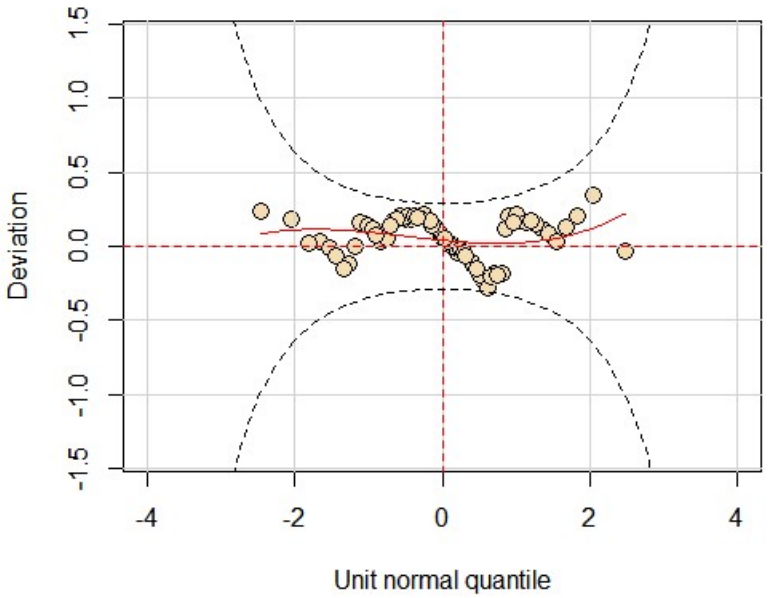
No entanto, nesse caso, recomenda-se utilizar o modelo final ($\eta_i = \beta_1 + \beta_2 \mathrm{E}_i + \beta_3 \mathrm{RV}_i$), com inclusão da covariável apenas no submodelo da mediana. Embora seja possível atribuir regressão a todos os parâmetros, recomenda-se que isso ocorra quando se deseja atribuir maior flexibilidade ao ajuste dos modelos, especialmente no tocante aos parâmetros ϕ, μ e c, que não possuem interpretação prática.

Tabela 5.5: Estimativas, erro padrões, estatística z e p-valores para o modelo de regressão BAR em todos os parâmetros.

	Estimativa	Erro padrão	Estatística z	p-valor
submodelo para μ_d				
β_1	1,116	0,126	8,895	$< 0,001$
β_2	0,187	0,045	4,185	$< 0,001$
β_3	-0,333	0,034	-9,785	$< 0,001$
submodelo para ϕ				
β_1	5,636	0,919	6,134	$< 0,001$
β_2	-0,234	0,253	-0,923	0,359
submodelo para μ				
β_1	0,446	0,018	25,111	$< 0,001$
β_2	0	0,005	-0,028	0,978
submodelo para c				
β_1	0,442	0,018	24,698	$< 0,001$
β_2	0,001	0,005	0,186	0,853

Fonte: Elaboração própria.

Figura 5.6: *Worm plot* dos resíduos quantílicos do modelo de regressão BAR em todos os parâmetros.



Fonte: Elaboração própria.

Capítulo 6

Conclusões

6.1 Considerações Finais

É notória a diversidade de aplicação da distribuição beta na literatura, especialmente para descrever e modelar dados contínuos no intervalo $(0, 1)$, que, usualmente, são taxas e proporções. A utilização é observada em diversos contextos, desde análises de variáveis socioeconômicas e demográficas, até ambientais e epidemiológicas. Em se tratando de dados ambientais, pode haver desafios para modelagem de variáveis climáticas, que podem constituir fenômenos complexos, e apresentar formas com características não convencionais, como a bimodalidade ou caudas pesadas.

Entre as soluções para lidar com dados bimodais, a literatura destaca: modelos de misturas de distribuições, que, em geral, envolvem adição de muitos parâmetros, podendo gerar estimativas não consistentes; e modelos cujos parâmetros são de difícil interpretação. Considerando esse panorama, o presente trabalho teve como objetivo desenvolver um novo modelo de regressão que permitisse modelar dados bimodais em um intervalo real, e que apresentasse interpretação simples da relação entre a variável resposta e variáveis explicativas.

Nesse sentido, este trabalho apresenta contribuições na área de modelos de regressão que estendem os modelos de regressão beta. Primeiro foi desenvolvida uma nova distribuição de

probabilidade baseada na distribuição beta e que comporta bimodalidade, denominada de Beta Ampliada Reparametrizada – BAR. A nova distribuição também acomoda dados unimodais e o seu suporte pode ir além do intervalo $(0, 1)$, acomodando, inclusive, os valores 0 e 1. Segundo, foi proposta uma classe de modelos de regressão associada à distribuição BAR, a qual possibilita a interpretação das relações entre as covariáveis e a mediana da variável resposta. Terceiro, foi realizada a implementação da função BAR considerando δ fixado em $0{,}5$, isto é, $\mathrm{BAR}(\mu, \phi, \mu_{\mathrm{d}}, \mathrm{c}, \delta = 0{,}5)$, na estrutura da classe GAMLSS para ajustes de modelos de regressão. Essa última contribuição viabiliza a aplicação do modelo proposto por pesquisadores, cabendo destacar que os códigos produzidos neste trabalho estão disponíveis em `https://github.com/talia499/BARRegression.git`. Quarto, foi conduzida uma aplicação do novo modelo a dados reais. Tais dados se referem a medições climáticas da cidade de Goiânia, no período que compreende os anos de 2011 a 2022, e a variável resposta de interesse foi umidade relativa mínima do ar (URA). O ajuste independente e identicamente distribuído de URA resultou na escolha da distribuição BAR para modelar URA, ao invés da distribuição beta. A regressão com inclusão de covariáveis também apresentou desempenho satisfatório, indicando que a estação e a rajada máxima de vento são significativas para explicar a mediana da umidade relativa mínima do ar. Essa relação ocorre de modo que a estação chuvosa apresenta maior URA mediana que a estação seca; para uma mesma estação, o aumenta na rajada máxima de vento provoca uma diminuição da URA mediana.

Ademais, é relevante destacar que o presente trabalho apresenta algumas limitações, em razão, principalmente, do caráter pioneiro do estudo e do tempo disponível para sua execução. Entre as limitações pode-se citar: a não obtenção de propriedades matemáticas da distribuição BAR; a utilização do tradicional método de máxima verossimilhança para obter as estimativas dos parâmetros dos modelos de regressão BAR propostos; a dependência dos valores iniciais do algoritmo para a convergência e adequação de ajustes dos modelos; e, a ausência de um estudo de simulação para ilustrar o desempenho do estimador de máxima verossimilhança sob o modelo de regressão BAR.

Diante do exposto, sugere-se os seguintes tópicos para aprofundamento em trabalhos futuros:

I) obtenção de propriedades matemáticas da distribuição BAR, por exemplo, os momentos, em especial, média e variância. Isso seria relevante para avaliar a possibilidade de interpretar o parâmetro ϕ em termos da precisão da distribuição, conforme os apontamentos da análise gráfica dos parâmetros desempenhada na Seção 3;

II) implementação de métodos alternativos à máxima verossimilhança para o modelo de regressão BAR, dada a da distribuição BAR.

III) desenvolvimento de um estudo de simulação para avaliar o desempenho do modelo de regressão BAR na estimação dos coeficientes de regressão, bem como dos erros-padrão associados. É interessante que a análise contemple diferentes cenários, iniciando pela modelagem apenas do parâmetro μ_d, até a inclusão progressiva de regressão em todos os parâmetros.

Referências Bibliográficas

Akaike, Hirotugu (1974). "A New Look at the Statistical Model Identification". *IEEE Transactions on Automatic Control* 19.6, pp. 716–723. DOI: 10.1109/TAC.1974.1100705.

Bernardo, J. S., Almeida, F. M. de e Nascimento, A. C. C. (2020). "General quality of municipal education and the influences of public spending". *Education Policy Analysis Archives* 28.7, p. 23. DOI: 10.14507/epaa.28.4696. URL: https://epaa.asu.edu/index.php/epaa/article/view/4696.

Burnham, Kenneth P. e Anderson, David R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. 2ª ed. New York: Springer. ISBN: 978-0-387-95364-9.

Buuren, S.v. e Fredriks, M. (2001). "Worm plot: a simple diagnostic device for modelling growth reference curves". *Statistics in Medicine* 20.8, pp. 1259–1277.

Cole, T. J. e Green, P. J. (1992). "Smoothing reference centile curves: The LMS method and penalized likelihood". *Statistics in Medicine* 11, pp. 1305–1319.

Cribari-Neto, F e Zeileis, A (2010). "Beta Regression in R". $journal of Statistical Software$ 34, pp. 1–24. DOI: 10.18637/jss.v034.i02<https://doi.org/10.18637/jss.v034.i02>..

Davison, A. C. (2003). *Statistical Models*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press.

Dunn, P.k. e Smyth, RL (1996). "Randomized Quantile Residuals". *Journal of Computational and Graphical Statistics* 5.3, pp. 236–244. ISSN: 10618600.

Ferrari, S. e Cribari-Neto, F. (2004). "Beta regression for modelling rates and proportions". *Journal of Applied Statistics* 31.7, 799–815.

Gupta, Arjun K. e Nadarajah, Saralees (2004). *Handbook of Beta Distribution and Its Applications*. Boca Raton, FL: CRC Press.

Heisterkamp, S.H. e Pennings, J.L.A. (2004). "The use of a finite mixture of beta distributions in the analysis of microarray data". *Kwantitatieve Methoden* ?

Ji, Yuan et al. (fev. de 2005). "Applications of beta-mixture models in bioinformatics". *Bioinformatics* 21.9, pp. 2118–2122. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bti318. eprint: https://academic.oup.com/bioinformatics/article-pdf/21/9/2118/48972716/bioinformatics_21_9_2118.pdf. URL: https://doi.org/10.1093/bioinformatics/bti318.

Jones, E. R. et al. (2022). "Indoor humidity levels and associations with reported symptoms in office buildings". *Indoor Air*.

Leite, Maysa de Lima e Virgens Filho, Jorim Sousa das (2013). "Avaliação da distribuição beta como modelo probabilístico para análise de dados de velocidade do vento para Ponta Grossa - PR". *Revista Científica de Engenharia e Geociências* 3.1, pp. 41–49. URL: https://revistas.uepg.br/index.php/exatas/article/view/879.

Lisboa, Mathews de Noronha Silveira (2024). "Modelo de regressão para valores extremos bimodais". Dissertação de Mestrado em Estatística. Brasília: Universidade de Brasília.

Majumdar, Koyel et al. (dez. de 2024). "A novel family of beta mixture models for the differential analysis of DNA methylation data: An application to prostate cancer". *PLOS ONE* 19.12, pp. 1–21. DOI: 10.1371/journal.pone.0314014. URL: https://doi.org/10.1371/journal.pone.0314014.

Migliorati, Sonia, Di Brisco, Agnese Maria e Ongaro, Andrea (2018). "A new regression model for bounded responses". *Bayesian Analysis* 13.3, 845—872.

Montgomery, Douglas C., Peck, Elizabeth A. e Vining, G. Geoffrey (2021). *Introduction to Linear Regression Analysis*. 5th. Hoboken, NJ: John Wiley & Sons, Inc., Hoboken, New Jersey. ISBN: 978-0-470-54281-1.

Nelder, John A. e Wedderburn, Robert W.M. (1972). "Generalized linear models". *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 135.3, pp. 370–384.

Otiniano, C.E.G. et al. (2023). "A bimodal model for extremes data". *Environmental and Ecological Statistics* 30, pp. 261–288. DOI: 10.1007/s10651-023-00566-7. URL: https://doi.org/10.1007/s10651-023-00566-7.

Ribeiro, T.K.A. e Ferrari, S.L.P (2023). "Robust estimation in beta regression via maximum L_q-likelihood". *Statistical Papers* 64, pp. 321–353.

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

Rigby, R.A. e Stasinopoulos, D.M. (2005). *Generalized additive models for location, scale and shape*. Chapman e Hall/CRC, pp. 507–554.

Rigby, Robert et al. (set. de 2019). *Distributions for Modeling Location, Scale, and Shape: Using GAMLSS in R*. ISBN: 9780367278847. DOI: 10.1201/9780429298547.

Schmidt, Alexandra M., Moraes, Caroline P. de e Migon, Helio S. (2015). "A hierarchical dynamic beta regression model of school performance in the brazilian mathematical olympiads for public schools". *arXiv preprint arXiv:1507.00565*. URL: https://arxiv.org/abs/1507.00565.

Schröder, C. e Rahmann, S. (2017). "A hybrid parameter estimation algorithm for beta mixtures and applications to methylation state classification". *Algorithms for molecular biology : AMB* 12.21. DOI: https://doi.org/10.1186/s13015-017-0112-1. URL: https://epaa.asu.edu/index.php/epaa/article/view/4696.

Silva, B. L. Simões e, Otiniano, C. E. G. e Nakano, E. Y. (2024). "The return period of heterogeneous climate data with a new invertible distribution". *Stochastic Environmental Research*

and Risk Assessment 38, pp. 2283–2296. DOI: 10.1007/s00477-024-02679-2. URL: https://doi.org/10.1007/s00477-024-02679-2.

Simas, Alexandre B., Barreto-Souza, Wagner e Rocha, Andréa V. (2010). "Improved estimators for a general class of beta regression models". *Computational Statistics Data Analysis* 54.2, pp. 348–366. URL: https://ideas.repec.org/a/eee/csdana/v54y2010i2p348-366.html.

Smithson, M. e Verkuilen, J. (2006). "A Better Lemon Squeezer? Maximum-Likelihood Regression With Beta-Distributed Dependent Variables". *Psychological Methods, American Psychological Association* 11.1, pp. 54–71.

Stasinopoulos, D. M. et al. (mar. de 2017). *Flexible Regression and Smoothing: Using GAMLSS in R*. Chapman e Hall/CRC. ISBN: 9781138197909. DOI: 10.1201/b21973.

Vila, R. e Niyazi Çankaya, M. (2021). "A bimodal Weibull distribution: properties and inference". *Journal of applied statistics* 49.12. PMID: 38476621, pp. 3044–3062. DOI: 10.1080/02664763.2022.2146661. eprint: https://doi.org/10.1080/02664763.2021.193182. URL: https://doi.org/10.1080/02664763.2021.193182.

Vila, Roberto et al. (2024). "A model for bimodal rates and proportions". *Journal of Applied Statistics* 51.4. PMID: 38476621, pp. 664–681. DOI: 10.1080/02664763.2022.2146661. eprint: https://doi.org/10.1080/02664763.2022.2146661. URL: https://doi.org/10.1080/02664763.2022.2146661.

Apêndice A

Prova de validade da função de ligação logit inversa

Considere a função

$$g : (-1, 0) \longrightarrow \mathbb{R}, \qquad g(\tau) = \log\left(\frac{-\tau}{1+\tau}\right).$$

Mostra-se a seguir que g é uma função de ligação válida para o parâmetro $\tau \in (-1, 0)$.

Definição. Note que para todo $\tau \in (-1, 0)$, tem-se $-\tau > 0$ e $1 + \tau > 0$. Assim,

$$\frac{-\tau}{1+\tau} > 0,$$

o que garante que $g(\tau)$ está bem definida, uma vez que o logaritmo natural é definido apenas para argumentos positivos.

Monotonicidade estrita. Veja que $g(\tau)$ pode ser escrita da seguinte forma:

$$g(\tau) = \log(-\tau) - \log(1 + \tau),$$

assim, a derivada de g é dada por

$$\begin{aligned} g'(\tau) &= \frac{d}{d\tau}\log(-\tau) - \frac{d}{d\tau}\log(1+\tau) \\ &= \frac{1}{\tau} - \frac{1}{1+\tau} \\ &= \frac{(1+\tau)-\tau}{\tau(1+\tau)} \\ &= \frac{1}{\tau(1+\tau)}. \end{aligned}$$

Para $\tau \in (-1, 0)$, tem-se $\tau < 0$ e $1 + \tau > 0$, de modo que $\tau(1 + \tau) < 0$, assim

$$g'(\tau) < 0, \qquad \forall\, \tau \in (-1, 0).$$

Logo, g é estritamente decrescente em $(-1, 0)$ e, portanto, injetiva.

Limites nos extremos do domínio. Quando $\tau \to 0^-$, tem-se $-\tau \to 0^+$ e $1+\tau \to 1$, de modo que

$$\frac{-\tau}{1+\tau} \to 0^+ \quad \Longrightarrow \quad \lim_{\tau\to 0^-} g(\tau) = \log(0^+) = -\infty.$$

Por outro lado, quando $\tau \to -1^+$, tem-se $-\tau \to 1$ e $1 + \tau \to 0^+$, de modo que

$$\frac{-\tau}{1+\tau} \to +\infty \quad \Longrightarrow \quad \lim_{\tau\to -1^+} g(\tau) = \log(+\infty) = +\infty.$$

Como g é contínua e estritamente monótona em $(-1, 0)$, segue que

$$g\big((-1, 0)\big) = \mathbb{R},$$

isto é, g é sobrejetiva sobre $\mathbb{R}$.

Função inversa. Seja $\eta = g(\tau)$. Então,

$$\eta = \log\left(\frac{-\tau}{1+\tau}\right),$$

o que implica em

$$e^{\eta} = \frac{-\tau}{1+\tau}.$$

Logo,

$$\begin{aligned} e^{\eta}(1+\tau) &= -\tau \\ e^{\eta} + e^{\eta}\tau &= -\tau \\ \tau(e^{\eta}+1) &= -e^{\eta} \\ \tau &= -\frac{e^{\eta}}{1+e^{\eta}} = g^{-1}(\eta). \end{aligned}$$

Observa-se que, para todo $\eta \in \mathbb{R}$,

$$0 < \frac{e^{\eta}}{1+e^{\eta}} < 1 \quad \Longrightarrow \quad -1 < \tau < 0,$$

garantindo que $g^{-1} : \mathbb{R} \to (-1, 0)$.

Conclusão. A função

$$g(\tau) = \log\left(\frac{-\tau}{1+\tau}\right)$$

é uma função de ligação válida para $\tau \in (-1, 0)$, pois é bem definida, estritamente monótona, sobrejetiva sobre $\mathbb{R}$ e possui inversa explícita dada por

$$g^{-1}(\eta) = -\frac{e^{\eta}}{1+e^{\eta}}.$$