Universidade de Brasília
Departamento de Estatística

Estudo sobre Modelos Aditivos Generalizados para Localização, Escala e Forma: Teoria e Aplicações sob Modelos Paramétricos

Stefan Zurman Gonçalves

Relatório apresentado para o Departamento de Estatística da Universidade de Brasília como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Brasília
2024

Stefan Zurman Gonçalves

Estudo sobre Modelos Aditivos Generalizados para Localização, Escala e Forma: Teoria e Aplicações sob Modelos Paramétricos

Orientadora: Profa. Terezinha Késsia de Assis Ribeiro

Relatório apresentado para o Departamento de Estatística da Universidade de Brasília como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Brasília
2024

Agradecimentos

Aos meus pais Davi e Zora por me darem todo o suporte necessário para chegar aqui, nunca poupando esforços para me ajudar, e celebrando todas minhas conquistas.

Ao meu irmão Robert por todas as boas memórias que tornaram minha vida melhor, e todo o apoio que tornou mais leves todos meus momentos difíceis.

Aos meus professores, dos quais todos os aprendizados me trouxeram até aqui.

Aos amigos que fiz durante minha trajetória, por compartilharem todos bons momentos comigo.

À minha orientadora Terezinha, por ter me guiado no trabalho, pelo apoio incansável, pelos ensinamentos enquanto minha professora, pelo apoio e incentivo na minha aplicação para cursos de pós-graduação, e pela ajuda na correria do final do semestre.

Por fim, agradeço à Fundação de Apoio a Pesquisa do Distrito Federal pelo financiamento do projeto de Iniciação Científica desenvolvido em conjunto com o atual projeto.

Resumo

O estudo de modelos de regressão tem sido um dos principais tópicos de estudo na estatística. Dentro desse contexto, os Modelos Aditivos Generalizados para Localização, Escala e Forma (GAMLSS) se destacam por sua flexibilidade. Esta característica decorre da capacidade dessa classe de modelos acomodar uma ampla variedade de distribuições de probabilidade para a variável resposta, e da possibilidade de ajustar estruturas de regressão para cada parâmetro da distribuição. Além disso, a estrutura dos GAMLSS facilita a interpretação dos parâmetros dos modelos, especialmente sob modelos paramétricos. O presente trabalho explora as vantagens e limitações dos GAMLSS na modelagem estatística. Também são comparadas suas vantagens em relação aos modelos lineares generalizados e modelos de regressão linear normal. Adicionalmente, são discutidas as diferentes distribuições que podem ser incorporadas nos GAMLSS e métodos para a criação de novas distribuições a partir de distribuições existentes. Neste trabalho são destacadas situações em que cada distribuição é apropriada para dados de diferentes naturezas, e como essas distribuições atuam no contexto da regressão pelos GAMLSS. Os métodos inferenciais sob os GAMLSS paramétricos são discutidos. Técnicas e procedimentos para a seleção de diferentes modelos são abordados. As principais técnicas de diagnóstico baseadas em resíduos quantílicos para essa classe de modelos são aqui definidas e ilustradas. Por fim, foram ajustados modelos de regressão para dados reais, demonstrando-se a aplicabilidade da classe GAMLSS, o que permite identificar tanto as suas vantagens quanto as suas limitações.

Palavras-chaves: Distribuições; GAMLSS; Modelos Aditivos Generalizados para Localização, Escala e Forma; Regressão Paramétrica; Resíduo Quantílico; *Worm plot*.

Lista de Tabelas

Lista de Figuras

Sumário

1 Introdução

A modelagem na estatística visa explicar o comportamento de uma ou mais variáveis de interesse, denominadas variáveis resposta, a partir do comportamento de outras variáveis chamadas variáveis explicativas. A partir do entendimento dessa relação é possível inferir sobre o comportamento das variáveis resposta a partir de informações provindas das variáveis explicativas (Freedman, 2009). Devido a sua vasta aplicabilidade em diversas áreas, como as ciências sociais, biológicas e na área de negócios (Kutner et al., 2005), a área de modelagem tem sido um dos principais tópicos de estudo na estatística.

Um dos métodos de modelagem mais simples é a regressão linear normal, referido aqui como RLN. Nesse modelo é possível entender o comportamento da média de uma variável resposta a partir dos valores fixados de suas variáveis explicativas (Draper; Smith, 1998). Entretanto, para os resultados do modelo RLN serem confiáveis, algumas premissas sobre o comportamento dos dados devem ser atendidas. Uma das principais limitações decorre da suposição de que a variável resposta segue uma distribuição normal com variância constante. Devido a distribuição normal possuir suporte na reta real, esta suposição implica que a variável resposta também deve assumir valores reais. Todavia, essa suposição pode não refletir o comportamento real da variável de interesse. Por exemplo, pode-se ter uma resposta referente a dados de contagem, que além de serem estritamente positivos, devem necessariamente apresentar valores inteiros. Para acomodar adequadamente uma variedade maior de dados, Nelder e Wedderburn (1972) propuseram a classe de modelos lineares generalizados, também conhecidos como MLGs.

A classe dos MLGs permite que a distribuição de probabilidades da resposta seja parte de um conjunto maior de família de distribuições, conhecida por família exponencial uniparamétrica de distribuições (Paula, 2023). A partir desta classe de modelos, é possível que variáveis com diversos comportamentos sejam modeladas adequadamente, tais como variáveis de contagem, variáveis estritamente positivas, variáveis binárias, variáveis restritas a um intervalo, ou até mesmo variáveis definidas na reta real, como na RLN. Outra vantagem dos MLGs deve-se a função de ligação, que relaciona as variáveis explicativas com a média da variável resposta. As funções de ligação permitem que a relação entre as variáveis explicativas e a resposta média assuma diversas formas, e podem também garantir que a média da distribuição esteja bem definida em seu espaço paramétrico. Por estas razões, se torna possível ter um melhor ajuste do modelo aos dados através dos MLGs. Entretanto, mesmo com as flexibilidades provenientes dos MLGs, ainda há casos em que esses não são capazes de descrever de forma adequada o comportamento dos da-

dos. Outras classes de modelos estatísticos foram propostos para aumentar a flexibilidade dos ajustes e a representatividade dos modelos sobre os dados, como os modelos aditivos generalizados (Hastie, 2017), os modelos lineares generalizados mistos (Breslow; Clayton, 1993), e os modelos aditivos generalizados de localização, escala e forma, também chamados de GAMLSS (Rigby; Stasinopoulos, 2005), que são o foco desse trabalho.

Como discutido por Rigby e Stasinopoulos (2005), a classe GAMLSS assume que a variável resposta pode seguir uma família de distribuições mais geral, podendo ter até 4 parâmetros. Essa classe de modelos permite que todos os 4 parâmetros sejam modelados através de estruturas de regressão. Assim como os MLGs, os GAMLSS incorporam a função de ligação nas estruturas de regressão associadas aos parâmetros. No entanto, os GAMLSS abrangem uma família de distribuições muito mais ampla do que a família exponencial uniparamétrica. Ademais, os GAMLSS podem acomodar uma estrutura mais complexa de como as variáveis explicativas são incorporadas dentro da função de ligação, não necessariamente através de uma relação linear com os parâmetros de regressão. Sendo assim, pode-se ver os modelos anteriormente mencionados como casos especiais dos GAMLSS, ou seja, a classe GAMLSS é uma generalização dos modelos citados até agora. Devido às flexibilidades viabilizadas pelos GAMLSS, será possível obter um ajuste do modelo muito mais fidedigno aos dados, possibilitando inferências mais precisas sobre o comportamento da variável resposta e sua relação com as variáveis explicativas.

O presente trabalho, que está dividido em 5 capítulos, visa o estudo e comparação entre o modelo de RLN, os MLGs e os GAMLSS. No Capítulo 2, é apresentado o referencial teórico deste trabalho. Nas Seções 2.1 e 2.2 são descritos os conceitos de família exponencial e família de localização-escala, que são necessários para introduzir os MLGs e GAMLSS. Nas Seções 2.3, 2.4, 2.5 são discutidos métodos para criação de novas distribuições. Nas Seções 2.6 a 2.10 são destacadas distribuições utilizadas na modelagem através dos GAMLSS que são adequadas para cada tipo de suporte da variável resposta. As Seções 2.11, 2.12 e 2.13 detalham os modelos de RLN, MLGs e GAMLSS, respectivamente. A Seção 2.14 aborda as técnicas inferenciais sob os GAMLSS. Na Seção 2.15 são explicadas ferramentas e critérios para seleção de modelos para essa classe de modelos. Na Seção 2.16 são fornecidas técnicas de diagnóstico para os GAMLSS. O Capítulo 3 relata a metodologia do estudo, além de detalhar os recursos computacionais e *softwares* utilizados neste trabalho. Aplicações a dados reais são ilustradas no Capítulo 4, onde há um foco nas vantagens sobre modelos usuais, e são mostrados como são feitos os ajustes sob esses modelos e suas interpretações. Finalmente, no Capítulo 5 são enunciadas as considerações finais deste trabalho.

2 Referencial Teórico

Neste trabalho, visa-se fazer uma revisão sobre os modelos aditivos generalizados de localização, escala e forma, introduzidos por Rigby e Stasinopoulos (2005). Para se ter um entendimento adequado desses modelos, inicialmente são introduzidos conceitos necessários para definição dos modelos de regressão estudados, tais como família exponencial e família de localização-escala. Em seguida, são apresentadas as principais distribuições de probabilidade assumidas para uma variável resposta. Na sequência, será feita uma fundamentação teórica com relação aos modelos de regressão linear normal e modelos lineares generalizados, além de suas aplicabilidades e limitações. Assim, o seguinte capítulo tem como objetivo introduzir os principais pontos referentes as classes de modelos RLN, MLGs e GAMLSS, a fim de ser possível então introduzir os GAMLSS dentro da adequada contextualização.

2.1 Família Exponencial

A família exponencial uniparamétrica de distribuições é um conjunto de distribuições caracterizadas por uma forma específica. Distribuições como normal, binomial, gama, Poisson e Gaussiana inversa pertencem a essa família. Stasinopoulos et al. (2017) definem uma distribuição como pertencente à família exponencial uniparamétrica quando é possível expressar a função densidade ou função de probabilidade da variável aleatória Y a partir de

$$f(y;\theta,\phi) = \exp\left[\frac{1}{\phi}\{y\theta - b(\theta)\} + c(y,\phi)\right],$$

em que

- θ é o parâmetro de canônico da distribuição tal que $\theta = \theta(\mu)$ é função de $\mu = \mathrm{E}(Y)$, e está associado com medidas de tendência central;
- $b(\theta)$ é uma função de θ tal que $b'(\theta) = \mathrm{E}(Y)$, com $b'(\theta) = \frac{\partial b(\theta)}{\partial \theta}$;
- $c(y,\phi)$ é uma função de y e ϕ;
- ϕ é o parâmetro de dispersão da distribuição, associado com medidas de dispersão, tal que $\mathrm{Var}(Y) = \phi b''(\theta) = \phi V(\mu)$ com $b''(\theta) = \frac{\partial^2 b(\theta)}{\partial \theta^2}$, sendo $V(\mu)$ denominada de função de variância.

2.2 Família de Localização-Escala

Uma variável aleatória contínua Y é dita ter uma distribuição pertencente à família de localização-escala F com parâmetro de localização μ e parâmetro de escala σ se a variável definida por (Stasinopoulos et al., 2017)

$$Z = \frac{Y - \mu}{\sigma}$$

possui função de distribuição acumulada que não depende dos parâmetros μ e σ satisfazendo

$$F_Y(y) = F_Z\left(\frac{y - \mu}{\sigma}\right),$$

e, consequentemente,

$$f_Y(y) = \frac{1}{\sigma} f_Z\left(\frac{y - \mu}{\sigma}\right),$$

em que

- $F_Y(\cdot)$ é a função de distribuição acumulada da variável Y;
- $F_Z(\cdot)$ é a função de distribuição acumulada da variável Z;
- $f_Y(\cdot)$ é a função densidade de probabilidade da variável Y;
- $f_Z(\cdot)$ é a função densidade de probabilidade da variável Z.

Portanto, a partir da função de distribuição acumulada ou da função densidade de probabilidade da variável Z, e dos valores de μ e σ, é possível definir a variável Y a partir de $Y = \mu + \sigma Z$.

Uma vantagem de se trabalhar com distribuições de localização-escala é que modelos ajustados sob essas são invariante para diferentes transformações de localização-escala da variável resposta, quando usando função de ligação identidade para modelar μ e função de ligação logarítmica para modelar σ[1]. Também, a forma da distribuição dados os parâmetros de forma não depende de μ ou σ, tornando a interpretação dos parâmetros mais simples (Rigby et al., 2019).

[1]Funções de ligação são usadas no contexto de regressão para associar o preditor linear com parâmetros da distribuição da variável resposta.

2.3 Método de Azzalini

Com a ideia de criar distribuições assimétricas a partir de distribuições simétricas, Azzalini (1985) propôs que a partir de uma variável contínua X_1 com distribuição simétrica e centrada em zero, e uma variável contínua X_2 também com distribuição simétrica e centrada em zero, pode-se definir uma nova variável aleatória Z contínua cuja função de densidade de probabilidades é definida por

$$f_Z(z) = 2f_{X_1}(z)F_{X_2}(\nu z), \tag{2.3.1}$$

com $\nu \in \mathbb{R}$, em que $f_{X_1}(\cdot)$ e $F_{X_2}(\cdot)$ são as respectivas função de densidade de X_1 e função de distribuição acumulada de X_2.

A ideia do método é aplicar pesos na função densidade de probabilidade de X_1 a partir da função de distribuição acumulada de X_2, criando assim uma distribuição com assimetria. A partir das distribuições geradas por esse método, pode-se modelar a assimetria da distribuição de Z a partir do parâmetro ν. A partir de Z com função densidade dada pela Expressão (2.3.1), se uma variável Y for definida tal que $Y = \mu + \sigma Z$, onde $F_Z(\cdot)$ não depende de μ ou σ, a variável Y possui distribuição pertencente a uma família de localização-escala, com parâmetro de assimetria ν. A partir dessa transformação, a função densidade de probabilidade de Y é dada por

$$f_Y(y) = \frac{2}{\sigma} f_{X_1}(z)F_{X_2}(\nu z)$$

com $z = (y - \mu)/\sigma$. A principal desvantagem de distribuições geradas por esse método é que suas funções de distribuição acumulada não possuem forma analítica, e devem então ser obtidas a partir de resultados numéricos. Neste trabalho, quando uma distribuição for gerada a partir desse método, essa será caracterizada pela Expressão (2.3.1).

2.4 Método de Junção/União

O método de junção/união consiste em criar uma nova variável aleatória contínua Y com distribuição assimétrica a partir de duas variáveis aleatórias contínuas X_1 e X_2 com distribuições simétricas centradas em $\mu \in \mathbb{R}$, de tal forma que a função de densidade de probabilidade de Y é definida por

$$f_Y(y) = \pi_1 f_{X_1}(y)\mathrm{I}(y < \mu) + \pi_2 f_{X_2}(y)\mathrm{I}(y \geq \mu)$$

em que $I(\cdot)$ é a função indicadora retornando valor 1 quando $(\cdot)$ for verdadeiro, e 0 caso contrário. Para garantir que $f_Y(y)$ seja uma função de densidade bem definida e contínua em $y = \mu$, π_1 e π_2 podem ser definidos de tal forma que a função de densidade de probabilidade fica definida como

$$f_Y(y) = \frac{2}{1+k}\left[f_{X_1}(y)\mathrm{I}(y < \mu) + k f_{X_2}(y)\mathrm{I}(y \geq \mu)\right],$$

com $k = f_{X_1}(\mu)/f_{X_2}(\mu)$.

Se X_1 e X_2 são definidas tais que $X_1 = \mu + \sigma Z/\nu$ e $X_2 = \mu + \sigma\nu Z$, $\sigma > 0$ e $\nu > 0$, em que Z é uma variável aleatória com distribuição simétrica centrada em zero, a função de densidade de probabilidade de Y pode ser definida como

$$f_Y(y) = \frac{2\nu}{\sigma(1+\nu^2)}\left(f_Z(\nu z)\mathrm{I}(y < \mu) + k f_Z(z/\nu)\mathrm{I}(y \geq \mu)\right), \tag{2.4.1}$$

em que $k = f_{X_1}(\mu)/f_{X_2}(\mu) = \nu^2$ e $z = (y-\mu)/\sigma$ (Fernandez; Steel, 1998). Quando a variável aleatória Z possui função de distribuição acumulada que não depende dos parâmetros μ e σ, a variável aleatória Y possuirá distribuição pertencente a uma família de localização-escala, com parâmetro de localização μ, parâmetro de escala σ, e ν será o parâmetro de assimetria. Neste trabalho, uma distribuição gerada por esse método será descrita pela Expressão (2.4.1).

2.5 Distribuições Inflacionadas e Ajustadas

Distribuições inflacionadas ou infladas são obtidas a partir da mistura de duas distribuições, uma com toda massa de probabilidade concentrada em um valor k e a outra uma distribuição qualquer de interesse D_1. Assim, uma variável Y segue uma distribuição $\mathrm{D}_1(\boldsymbol{\theta}, p)$ inflacionada em k se $Y = k$ com probabilidade p e $Y \sim \mathrm{D}_1(\boldsymbol{\theta})$ com probabilidade $1 - p$. Essas distribuições recebem tal nome pois aumentam a probabilidade de $Y = k$, tal que a variável pode assumir valor k tanto devido à distribuição de interesse D_1 quanto devido à distribuição com massa de probabilidade concentrada no valor k. Geralmente as distribuições inflacionadas em k são utilizadas quando mais de um evento levam à uma variável assumir valor igual a k.

Distribuições ajustadas são similares às distribuições inflacionadas, de modo que essas são obtidas a partir da mistura de duas distribuições, uma com toda massa de probabilidade concentrada em um valor k e a outra uma distribuição qualquer de interesse D_1 truncada em k. Assim, uma variável Y segue uma distribuição $\mathrm{D}_1(\boldsymbol{\theta}, p)$ ajustada em

k se $Y = k$ com probabilidade p, e $Y = y \neq k$ com probabilidade ou densidade de probabilidade $(1-p)\frac{f_1(y)}{1-f_1(0)}$, onde $f_1(\cdot)$ é a função densidade de probabilidade da distribuição $\mathrm{D}_1(\boldsymbol{\theta})$. Portanto, a distribuição assume valor k apenas pelo componente da distribuição com massa de probabilidade concentrada no valor k, podendo essa probabilidade ser maior ou menor que a probabilidade de assumir valor k na distribuição original $\mathrm{D}_1(\boldsymbol{\theta})$. As distribuições ajustadas em k podem ser utilizadas para modelar variáveis quando algum evento desconhecido causa um acréscimo ou decréscimo de observações com valor igual a k.

Os tipos de distribuições citados anteriormente são usualmente utilizados quando um valor para a variável resposta nos dados apresenta frequência divergente do esperado segundo o modelo postulado, seja por um excesso desses valores ou por uma falta desses. Esse comportamento pode ocorrer no caso de dados discretos, se uma grande ou pequena parte dos dados estiver concentrada no valor zero, situação que recebe o nome de excesso ou falta de zeros. Também, pode ocorrer de dados na teoria contidos entre (0,1) apresentarem valores iguais a zero ou iguais a um, principalmente quando a variável resposta medir alguma proporção ou taxa. Nesses casos, distribuições inflacionadas em zero ou um devem ser utilizadas, uma vez que distribuições contínuas atribuem probabilidade zero para dados fora do intervalo (0,1), e então a modelagem por uma distribuição contínua levaria a um ajuste inadequado. Pode ocorrer também de variáveis contínuas estritamente positivas apresentarem valores iguais a zero, onde também são necessárias as distribuições inflacionadas em zero para o ajuste correto aos dados.

2.6 Distribuições Contínuas na Reta Real

A escolha de uma distribuição adequada para a variável resposta é uma das etapas mais importantes na especificação de modelos de regressão, sendo crucial o entendimento das propriedades da variável resposta em cada banco de dados para esta escolha (Stasinopoulos et al., 2024). Assim, o entendimento das distribuições disponíveis sob os GAMLSS é crucial para o ajuste destes modelos. A Tabela 1 apresenta todas as distribuições contínuas na reta real atualmente presentes no pacote *gamlss* do R, além de suas abreviações e respectivos parâmetros. O pacote *gamlss* do R é utilizado para ajustar os modelos que serão definidos na Seção 2.13. Estas distribuições são utilizadas majoritariamente quando a variável resposta pode assumir valores tanto positivos quanto negativos não sendo restritos a nenhum intervalo, e apresentam valores contínuos. Entretanto, Stasinopoulos et al. (2024) sugerem que na prática essas distribuições podem ser utilizadas para modelar dados estritamente positivos ou limitados ao intervalo $(0, 1)$, caso

a variável resposta tenha valores observados suficientemente distantes desses limites tal que os valores ajustados não estejam fora dos limites.

Tabela 1: Distribuições contínuas na reta real e suas siglas.

Distribuição	Sigla	Distribuição	Sigla
Gaussiana-Exponencial	exGAUS(μ, σ, ν)	*Sinh-arcsinh* Original Reparametrizada	SHASHo2(μ, σ, ν, τ)
Beta Exponencial Generalizada Tipo 2	EGB2(μ, σ, ν, τ)	Normal Assimétrica Tipo 1	SN1(μ, σ, ν)
T Generalizada	GT(μ, σ, ν, τ)	Normal Assimétrica Tipo 2	SN2(μ, σ, ν)
Gumbel	GU(μ, σ)	*Exponential Power* Assimétrica Tipo 1	SEP1(μ, σ, ν, τ)
SU de Johnson Original	JSUo(μ, σ, ν, τ)	*Exponential Power* Assimétrica Tipo 2	SEP2(μ, σ, ν, τ)
SU de Johnson Reparametrizada	JSU(μ, σ, ν, τ)	*Exponential Power* Assimétrica Tipo 3	SEP3(μ, σ, ν, τ)
Logistica	LO(μ, σ)	*Exponential Power* Assimétrica Tipo 4	SEP4(μ, σ, ν, τ)
Normal-Exponencial-T	NET(μ, σ, ν, τ)	T Assimétrica Tipo 1	ST1(μ, σ, ν, τ)
Normal	NO(μ, σ)	T Assimétrica Tipo 2	ST2(μ, σ, ν, τ)
Normal Reparametrizada	NO2(μ, σ)	T Assimétrica Tipo 3	ST3(μ, σ, ν, τ)
Família Normal	NOF(μ, σ, ν)	T Assimétrica Tipo 3 Reparametrizada	SST(μ, σ, ν, τ)
Power Exponential	PE(μ, σ, ν)	T Assimétrica Tipo 4	ST4(μ, σ, ν, τ)
Power Exponential Reparametrizada	PE2(μ, σ, ν)	T Assimétrica Tipo 5	ST5(μ, σ, ν, τ)
Gumbel Reversa	RG(μ, σ)	Família T	TF(μ, σ, ν)
Sinh-arcsinh	SHASH(μ, σ, ν, τ)	Família T Tipo 2	TF2(μ, σ, ν)
Sinh-arcsinh Original	SHASHo(μ, σ, ν, τ)		

Das distribuições indicadas na Tabela 1, as distribuições SN1, SEP1, SEP2, ST1 e ST2 são obtidas a partir do método de Azzalini, e as distribuições SN2, SEP3, SEP4, ST3 e ST4 são obtidas pelo método de junção/união. Das distribuições apresentadas, com a exceção das distribuições exGAUS, NO2, e NOF, todas pertencem à família de localização-escala, o que torna os parâmetros das distribuições e os coeficientes ajustados em modelos de regressão de fácil interpretação. Entre as principais adversidades em modelar dados a partir de distribuições com suporte real está em encontrar uma distribuição que modele adequadamente a simetria e a curtose da variável resposta. A seguir são detalhadas algumas das distribuições da Tabela 1 comumente utilizadas para modelagem, e algumas que possuem flexibilidade para tais adversidades.

2.6.1 Normal

A distribuição normal é uma distribuição de localização-escala, com função densidade de probabilidade dada por

$$f(y|\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\},$$

em que $y \in \mathbb{R}$, $\mu \in \mathbb{R}$ é o parâmetro de localização e a média, e $\sigma > 0$ é o parâmetro de escala e desvio padrão. Se Y for uma variável aleatória com distribuição normal, denota-se

$Y \sim \text{NO}(\mu, \sigma)$. A sua função densidade de probabilidade pode ser reescrita na forma

$$f(y|\mu,\sigma) = \exp\left\{\frac{1}{\sigma^2}\left(y\mu - \frac{\mu^2}{2}\right) - \frac{1}{2}\left(\log(2\pi\sigma^2) + \frac{y^2}{\sigma^2}\right)\right\}.$$

Assim, a distribuição normal pertence à família exponencial uniparamétrica com

- $\theta = \mu$;
- $b(\theta) = \frac{\theta^2}{2}$, e $\text{E}(Y) = b'(\theta) = \mu$;
- $\phi = \sigma^2$, e $\text{Var}(Y) = \phi b''(\theta) = \sigma^2$;
- $c(y, \phi) = -\frac{1}{2}\left(\log(2\pi\sigma^2) + \frac{y^2}{\sigma^2}\right)$.

No contexto de modelagem, essa distribuição é utilizada para modelar variáveis resposta contidas em toda reta real, em que tanto nos modelos de regressão linear normal quanto nos MLGs, modela-se μ a partir das covariáveis. Também, para ambos RLN e MLG usual, assume-se que σ é constante para todos valores das covariáveis. Em contrapartida, para os GAMLSS, será possível não só modelar μ, mas também σ a partir das covariáveis. Usualmente, utiliza-se a função de ligação identidade quando modelando μ para os MLGs e GAMLSS, e função de ligação logarítmica para modelar σ nos GAMLSS. A distribuição normal não é adequada para modelagem quando a variável resposta dada as covariáveis não é definida em toda reta real, possui alguma assimetria ou apresenta comportamento leptocúrtico ou platicúrtico (Rigby et al., 2019).

2.6.2 Gumbel

A distribuição Gumbel é uma distribuição de localização-escala, com função densidade de probabilidade dada por (Johnson; Kotz; Balakrishnan, 1995)

$$f(y|\mu,\sigma) = \frac{1}{\sigma}\exp\left\{\left(\frac{y-\mu}{\sigma}\right) - \exp\left(\frac{y-\mu}{\sigma}\right)\right\},$$

em que $y \in \mathbb{R}$, $\mu \in \mathbb{R}$ é o parâmetro de localização e também a moda da distribuição, e $\sigma > 0$ é o parâmetro de escala. Se uma variável aleatória Y seguir distribuição Gumbel, denota-se $Y \sim \text{GU}(\mu, \sigma)$.

A distribuição Gumbel não pertence à família exponencial uniparamétrica, então não pode ser modelada por meio de MLG. Entretanto, sua modelagem é possível a partir dos GAMLSS. Sob estes, utiliza-se a função de ligação identidade para modelar μ e

ligação logarítmica para modelar σ. Essa distribuição possui assimetria à esquerda e é leptocúrtica, com valores para a assimetria e curtose fixos, independentes dos valores de μ e σ. Assim, a distribuição Gumbel é utilizada para dados com assimetria à esquerda e leptocúrticos, mas quando sua assimetria e curtose não sofrerem alterações devido a variações dos valores das covariáveis (Rigby et al., 2019).

2.6.3 Gumbel Reversa

A distribuição Gumbel reversa é uma distribuição de localização-escala, com função de densidade de probabilidade dada por (Johnson; Kotz; Balakrishnan, 1995)

$$f(y|\mu,\sigma) = \frac{1}{\sigma}\exp\left\{-\left(\frac{y-\mu}{\sigma}\right) - \exp\left[-\left(\frac{y-\mu}{\sigma}\right)\right]\right\},$$

em que $y \in \mathbb{R}$, $\mu \in \mathbb{R}$ é o parâmetro de localização e também a moda da distribuição, e $\sigma > 0$ é o parâmetro de escala. Denota-se $Y \sim \text{RG}(\mu,\sigma)$ quando uma variável aleatória Y segue distribuição Gumbel reversa.

A distribuição Gumbel reversa possui uma relação com a distribuição Gumbel de forma que se $Y \sim \text{GU}(\mu,\sigma)$, então $-Y \sim \text{RG}(-\mu,\sigma)$. Analogamente, essa distribuição não pertence à família exponencial uniparamétrica. Entretanto, sua modelagem é possível sob os GAMLSS. Sob esses, utiliza-se a função de ligação identidade para modelar μ e ligação logarítmica para modelar σ. Essa distribuição possui propriedades similares à distribuição Gumbel, de forma que essa possui assimetria à direita e é leptocúrtica, com valores para a assimetria e curtose fixos, independentes dos valores de μ e σ. Assim, essa distribuição é utilizada para dados com assimetria à direita e leptocúrticos, mas quando sua assimetria e curtose não sofrerem alterações devido a variações dos valores das covariáveis (Rigby et al., 2019)

2.6.4 *Power Exponential*

A distribuição *power exponential* é uma distribuição de localização-escala, com função de densidade de probabilidade dada por (Johnson; Kotz; Balakrishnan, 1995)

$$f(y|\mu,\sigma,\nu) = \frac{\nu\exp\left(-|z|^{\nu}\right)}{2c\Gamma(1/\nu)},$$

com $z = (y-\mu)/(c\sigma)$ e $c^2 = \Gamma(1/\nu)(\Gamma(3/\nu))^{-1}$, $y \in \mathbb{R}$, $\mu \in \mathbb{R}$ é o parâmetro de localização e também média, mediana e moda da distribuição; $\sigma > 0$ é o parâmetro de escala; e $\nu > 0$

é o parâmetro de forma, relacionado à curtose, com $\nu > 0$. Se Y segue uma distribuição *power exponential*, denota-se $Y \sim \text{PE}(\mu, \sigma, \nu)$.

A distribuição *power exponential* é simétrica e possui média dependente apenas de μ, variância dependente apenas de σ, e curtose dependente apenas de ν. Assim, a partir dessa distribuição é possível modelar de forma independente a média, a variância e a curtose da distribuição. Nos GAMLSS, utiliza-se a ligação identidade para μ, e ligação logarítmica para σ e ν. Casos especiais da distribuição *power exponential* são a distribuição de Laplace, quando $\nu = 1$, e a distribuição normal, quando $\nu = 2$. A distribuição *power exponential* pode ter comportamento leptocúrtico, mesocúrtico ou até platicúrtico, dependendo do valor de ν. Por fim, essa distribuição é apropriada para modelar dados simétricos com variações na curtose para diferentes valores das covariáveis (Rigby et al., 2019).

2.6.5 Normal Assimétrica Tipo 1

A distribuição normal assimétrica tipo 1 é uma distribuição de localização-escala, obtida a partir do Método de Azzalini, quando X_1 e X_2 seguem distribuição $\text{NO}(0, 1)$. Sua função de densidade de probabilidade é dada por (Azzalini, 1986)

$$f(y|\mu, \sigma, \nu) = \frac{2}{\sigma}\phi(z)\Phi(\nu z)$$

em que $y \in \mathbb{R}$, $z = (y - \mu)/\sigma$; $\mu \in \mathbb{R}$ é o parâmetro de localização com efeito aditivo na média; $\sigma > 0$ é o parâmetro de escala; $\nu \in \mathbb{R}$ é o parâmetro de forma relacionado à assimetria; $\phi(\cdot)$ é a função de densidade de probabilidade de uma distribuição $\text{NO}(0, 1)$; $\Phi(\cdot)$ é a função de distribuição acumulada da distribuição $\text{NO}(0, 1)$. Denota-se $Y \sim \text{SN1}(\mu, \sigma, \nu)$ quando a variável aleatória Y segue uma distribuição normal assimétrica tipo 1.

A partir dessa distribuição, pode-se modelar a simetria da distribuição a partir das covariáveis, além da localização e da escala. A distribuição SN1 possui assimetria à direita quando $\nu < 0$ e assimetria à esquerda se $\nu > 0$. Quando $\nu = 0$, a distribuição equivale a uma distribuição $\text{NO}(\mu, \sigma)$. Como a distribuição SN1 permite assimetria à esquerda e à direita, pode ser que dado os valores das covariáveis, a variável resposta seja assimétrica à esquerda, e para outros valores fixados das covariáveis, seja assimétrica à direita. Sob essa distribuição, a assimetria independe dos valores de μ e σ, dependendo apenas de ν. Além disso, a distribuição SN1 é também leptocúrtica. Nos GAMLSS, utiliza-se a ligação identidade para μ e ν, e ligação logarítmica para σ. A distribuição SN1 é apropriada para modelar dados leptocúrticos com diferenças na simetria da variável

resposta para diferentes valores das covariáveis (Rigby et al., 2019).

2.6.6 *Exponential Power* Assimétrica Tipo 1

A distribuição *exponential power* assimétrica tipo 1 é uma distribuição de localização-escala, obtida a partir do Método de Azzalini, quando X_1 e X_2 seguem distribuição $\text{PE}(0, \tau^{1/\tau}/c, \tau)$. Sua função de densidade de probabilidade é dada por (Azzalini, 1986)

$$f(y|\mu, \sigma, \nu, \tau) = \frac{2}{\sigma} f_{X_1}(z) F_{X_2}(\nu z),$$

em que $y \in \mathbb{R}$, $z = (y - \mu)/\sigma$; $\mu \in \mathbb{R}$ é o parâmetro de localização com efeito aditivo na média; $\sigma > 0$ é o parâmetro de escala; $\nu \in \mathbb{R}$ é um parâmetro de forma relacionado à assimetria; $\tau > 0$ é um parâmetro de forma relacionado à curtose; $f_{X1}(\cdot)$ é a função de densidade de probabilidade de uma distribuição $\text{PE}(0, \tau^{1/\tau}/c, \tau)$; $F_{X_2}(\cdot)$ é a função de distribuição acumulada de uma distribuição $\text{PE}(0, \tau^{1/\tau}/c, \tau)$. Denota-se $Y \sim \text{SEP1}(\mu, \sigma, \nu, \tau)$ quando a variável aleatória Y segue uma distribuição *exponential power* assimétrica tipo 1.

Utilizando essa distribuição, torna-se possível modelar a simetria, curtose, localização e escala da distribuição da variável resposta a partir das covariáveis. Quando $\nu = 0$, a distribuição equivale a uma distribuição *power exponential*. A distribuição SEP1 consegue assumir comportamento platicúrtico, mesocúrtico e leptocúrtico, além de poder apresentar assimetria à direita, assimetria à esquerda ou simetria. Sendo assim, essa distribuição é apropriada para modelar dados com diferentes comportamentos na curtose e na simetria para diferentes valores das covariáveis. Nos GAMLSS, utiliza-se a ligação identidade para μ e ν, e ligação logarítmica para σ e τ (Rigby et al., 2019).

2.7 Distribuições Contínuas Positivas

A Tabela 2 apresenta todas as distribuições contínuas positivas presentes no pacote *gamlss* do R, além de suas abreviações e respectivos parâmetros. Essas distribuições são utilizadas quando a variável resposta pode assumir valores estritamente positivos e contínuos.

Tabela 2: Distribuições contínuas estritamente positivas e suas siglas.

Distribuição	Sigla	Distribuição	Sigla
Box-Cox Cole e Green	BCCG(μ, σ, ν)	Gama Inversa	IGAMMA(μ, σ)
Box-Cox Cole e Green Original	BCCGo(μ, σ, ν)	Gaussiana Inversa	IG(μ, σ)
Box-Cox *Power Exponential*	BCPE(μ, σ, ν, τ)	Log-Normal	LOGNO(μ, σ)
Box-Cox *Power Exponential* Original	BCPEo(μ, σ, ν, τ)	Log-Normal Reparametrizada	LOGNO2(μ, σ)
Box-Cox T	BCT(μ, σ, ν, τ)	Família Log-Normal	LNO(μ, σ, ν)
Box-Cox T Original	BCTo(μ, σ, ν, τ)	Pareto Tipo 1	PARETO1o(μ, σ)
Exponencial	EXP(μ)	Pareto Tipo 2	PARETO2(μ, σ)
Gama	GA(μ, σ)	Pareto Tipo 2 Original	PARETO2o(μ, σ)
Família Gama	GAF(μ, σ, ν)	Weibull	WEI(μ, σ)
Beta Generalizada Tipo 2	GB2(μ, σ, ν, τ)	Weibull Reparametrizada 1	WEI2(μ, σ)
Gama Generalizada	GG(μ, σ, ν)	Weibull Reparametrizada 2	WEI3(μ, σ)
Gaussiana Inversa Generalizada	GIG(μ, σ, ν)		

Entre as distribuições de probabilidade indicadas na Tabela 2 com a exceção das distribuições GAF, IG, LNO, LOGNO e WEI2, todas possuem parâmetro de escala, o que facilita a interpretação de tais parâmetros através dos coeficientes ajustados via modelos de regressão. Algumas adversidades quando modelando dados através de distribuições com suporte positivo são encontrar uma distribuição que modele adequadamente a simetria e a curtose da variável resposta, e a sua forma de maneira geral, além da dificuldade de modelar valores muito próximos a zero da variável resposta. A seguir são detalhadas algumas das distribuições indicadas na Tabela 2 que são comumente utilizadas para modelagem, e algumas que possuem flexibilidade para as adversidades citadas.

2.7.1 Exponencial

A distribuição exponencial é uma distribuição contínua estritamente positiva, e a única distribuição contínua com a propriedade de falta de memória. Sua função de densidade de probabilidade é dada por (Johnson; Kotz; Balakrishnan, 1994)

$$f(y|\mu) = \frac{1}{\mu} \exp\left(-\frac{y}{\mu}\right),$$

em que $y > 0$, $\mu > 0$ é a média, e também parâmetro de escala. Quando uma variável aleatória Y segue distribuição exponencial, denota-se $Y \sim$EXP(μ). A sua função densidade de probabilidade pode ser reescrita na forma

$$f(y|\mu) = \exp\left\{-y/\mu - \log(\mu)\right\}.$$

Assim, a distribuição exponencial pertence à família exponencial uniparamétrica, com

- $\theta = -1/\mu$;
- $b(\theta) = \log(-\theta)$, e $\mathrm{E}(Y) = b'(\theta) = \mu$;
- $\phi = 1$, e $\mathrm{Var}(Y) = \phi b''(\theta) = \mu^2$;
- $c(y, \phi) = 0$.

A distribuição exponencial é um caso especial da distribuição gama e da distribuição Weibull. Essa é apropriada para modelar dados com assimetria à direita, com densidade de probabilidade estritamente decrescente (moda tendendo a 0), e com a propriedade de perda de memória. Nos MLG e GAMLSS, modela-se μ a partir das covariáveis utilizando uma função de ligação logarítmica. Uma vez que a distribuição é uni-paramétrica, não é possível modelar a variância da variável resposta independentemente da sua média, e tampouco é possível modelar a assimetria, curtose ou forma da distribuição, sendo nesses casos necessárias distribuições com mais parâmetros (Rigby et al., 2019).

2.7.2 Gama

A distribuição gama é uma distribuição contínua estritamente positiva, com função de densidade de probabilidade dada por (Johnson; Kotz; Balakrishnan, 1994)

$$f(y|\mu, \sigma) = \frac{y^{1/\sigma^2}}{(\sigma^2\mu)^{1/\sigma^2} y \Gamma(1/\sigma^2)} \exp\left(-\frac{y}{\sigma^2\mu}\right),$$

em que $y > 0$, $\mu > 0$ é a média, e também parâmetro de escala, e $\sigma > 0$ é o coeficiente de variação, e também parâmetro que afeta assimetria e curtose. Quando uma variável aleatória Y segue distribuição gama, denota-se $Y \sim \mathrm{GA}(\mu, \sigma)$. A sua função densidade de probabilidade pode ser reescrita na forma

$$f(y|\mu, \sigma) = \exp\left\{\frac{1}{\sigma^2}[-y/\mu - \log(\mu)] + \frac{1}{\sigma^2}\log(y/\sigma^2) - \log(y) - \log(\Gamma(1/\sigma^2))\right\}.$$

Assim, a distribuição gama pertence à família exponencial uniparamétrica, com

- $\theta = -1/\mu$;
- $b(\theta) = \log(-\theta)$, e $\mathrm{E}(Y) = b'(\theta) = \mu$;
- $\phi = \sigma^2$, e $\mathrm{Var}(Y) = \phi b''(\theta) = \mu^2\sigma^2$;

- $c(y, \phi) = \frac{1}{\sigma^2}\log(y/\sigma^2) - \log(y) - \log(\Gamma(1/\sigma^2))$.

No contexto de modelagem, a distribuição gama é utilizada para modelar variáveis resposta contínuas estritamente positiva, com assimetria positiva e leptocúrticas. Nos MLG e GAMLSS, modela-se μ a partir das covariáveis, sendo μ a média da variável resposta, utilizando uma função de ligação inversa ou uma função de ligação logarítmica. Nos MLGs, tem-se a peculiaridade da variância da variável resposta aumentar conforme a média aumenta, uma vez que não há estrutura de regressão para σ. Nos GAMLSS, é possível modelar diretamente o coeficiente de variação a partir de σ, onde utiliza-se normalmente a função de ligação logarítmica. Assim, pode-se modelar a escala da distribuição a partir das covariáveis. A distribuição gama é extremamente útil para modelar dados estritamente positivos, principalmente na presença de assimetria nos dados. Entretanto, não é possível modelar diretamente a assimetria sob essa distribuição, e essa também não é adequada para dados platicúrticos (Rigby et al., 2019).

2.7.3 Gaussiana Inversa

A distribuição Gaussiana inversa é uma distribuição contínua estritamente positiva, com função de densidade de probabilidade dada por (Johnson; Kotz; Balakrishnan, 1994)

$$f(y|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2 y^3}} \exp\left(-\frac{1}{2\mu^2\sigma^2 y}(y-\mu)^2\right),$$

em que $y > 0$, $\mu > 0$ é a média, além de ter efeito na variância, assimetria e curtose, e $\sigma > 0$ possui efeito multiplicativo no desvio padrão, além de efeito na assimetria e curtose. Quando uma variável aleatória Y segue distribuição Gaussiana inversa, denota-se $Y \sim \text{IG}(\mu, \sigma)$. A sua função de densidade pode ser reescrita na forma

$$f(y|\mu, \sigma) = \exp\left\{\frac{1}{\sigma^2}\left(-\frac{y}{2\mu^2} + \frac{1}{\mu}\right) - \frac{1}{2}\left(\log(2\pi\sigma^2 y^3) + \frac{1}{\sigma^2 y}\right)\right\}.$$

Assim, a distribuição Gaussiana inversa pertence à família exponencial uniparamétrica, com

- $\theta = -1/(2\mu^2)$;
- $b(\theta) = -(-2\theta)^{1/2}$, e $\text{E}(Y) = b'(\theta) = \mu$;
- $\phi = \sigma^2$, e $\text{Var}(Y) = \phi b''(\theta) = \mu^3\sigma^2$;

- $c(y, \phi) = -\frac{1}{2}\left(\log(2\pi\sigma^2 y^3) + \frac{1}{\sigma^2 y}\right)$.

Na modelagem, a distribuição Gaussiana inversa é utilizada em cenários similares à distribuição gama. Entretanto, comparada com a distribuição gama, a distribuição Gaussiana inversa possui caudas mais pesadas, e assimetria maior para os mesmos valores fixados de média e variância. Também, na distribuição Gaussiana inversa, a variância depende de um termo cúbico da média, enquanto na gama depende de um termo quadrático, ou seja, a variância cresce mais rapidamente na distribuição Gaussiana inversa conforme a média aumenta. Nos MLG e GAMLSS, modela-se μ a partir de uma função de ligação logarítmica e raramente com uma função de ligação inversa quadrática. Nos GAMLSS, é possível modelar σ, em que utiliza-se normalmente a função de ligação logarítmica. A distribuição IG também é útil para modelar dados estritamente positivos, principalmente na presença de assimetria forte ou caudas pesadas. Entretanto, as mesmas limitações da distribuição gama se aplicam a essa distribuição, em que os parâmetros μ e σ que conjuntamente modelam a média e a variância da distribuição, também definem a assimetria e a curtose, além da inadequabilidade para dados com assimetria à esquerda, dados platicúrticos, ou dados homocedásticos (Rigby et al., 2019).

2.7.4 Gama Generalizada

A distribuição gama generalizada é uma distribuição contínua estritamente positiva, com três parâmetros, e com função de densidade de probabilidade dada por (Rigby et al., 2019)

$$f(y|\mu, \sigma, \nu) = \frac{|\nu|\theta^\theta z^\theta \exp(-\theta z)}{\Gamma(\theta) y},$$

em que $y > 0$, $\mu > 0$ é um parâmetro de escala; $\sigma > 0$ é um parâmetro de forma; $\nu \in \mathbb{R}$, $\nu \neq 0$ é um parâmetro de forma; $z = (y/\mu)^\nu$; $\theta = 1/(\sigma^2\nu^2)$. Se Y segue uma distribuição gama generalizada, denota-se $Y \sim \text{GG}(\mu, \sigma, \nu)$.

A distribuição GG permite ajustes mais flexíveis a dados contínuos positivos por possuir mais parâmetros de forma. As distribuições Weibull e gama são casos especiais da distribuição gama generalizada. Nos GAMLSS, utiliza-se a função de ligação logarítmica para μ e σ, e identidade para ν. Essa distribuição pode assumir forma com moda positiva ou moda tendendo a zero, e pode fornecer bons ajustes para dados onde a variável resposta apresenta valores muito próximos a zero (Rigby et al., 2019).

2.7.5 Box-Cox Cole e Green

A distribuição Box-Cox Cole e Green é uma distribuição contínua estritamente positiva, com três parâmetros, e com função de densidade de probabilidade dada por (Rigby; Stasinopoulos, 2004)

$$f(y|\mu,\sigma,\nu) = \frac{y^{\nu-1}\exp(-\frac{1}{2}z^2)}{\mu^{\nu}\sigma\sqrt{2\pi}\Phi\left[(\sigma|\nu|)^{-1}\right]},$$

em que $y > 0$, $\mu > 0$ é o parâmetro de escala e aproximadamente a mediana; $\sigma > 0$ é aproximadamente o coeficiente de variação; $\nu \in \mathbb{R}$ é um parâmetro de forma, relacionado à assimetria; $z = 1/(\sigma\nu)\left[(y/\mu)^{\nu} - 1\right]$, se $\nu \neq 0$, e $z = 1/(\sigma)\log(y/\mu)$, se $\nu = 0$; $\Phi(\cdot)$ é a função de distribuição acumulada da distribuição NO$(0,1)$. Denota-se $Y \sim$ BCCG(μ,σ,ν) quando a variável aleatória Y segue uma distribuição Box-Cox Cole e Green.

Utilizando essa distribuição, é possível modelar a simetria, localização e escala da distribuição da variável resposta a partir das covariáveis. A distribuição Box-Cox Cole e Green é um caso especial da distribuição Box-Cox *power exponential*, quando $\tau = 2$. A distribuição log-normal é um caso particular da distribuição Box-Cox Cole e Green. A distribuição BCCG consegue assumir comportamento assimétrico à direita e à esquerda, sendo então apropriada para modelar dados com diferentes comportamentos na simetria com variações nas covariáveis. Entretanto, a distribuição BCCG não consegue modelar a curtose de forma independente dos demais parâmetros da distribuição. Nos GAMLSS, os parâmetros μ e ν são modelados pelas covariáveis através da função de ligação identidade e o parâmetro σ é modelado a partir da função de ligação logarítmica. Vale notar que essa distribuição apresenta melhores ajustes para valores da variável resposta distantes de zero, e deve ser utilizada em tais casos (Rigby et al., 2019), apresentando geralmente moda positiva.

2.7.6 Box-Cox *power exponencial*

A distribuição Box-Cox *power exponencial* é uma distribuição contínua estritamente positiva, com quatro parâmetros, e com função de densidade de probabilidade dada por (Rigby; Stasinopoulos, 2004)

$$f(y|\mu,\sigma,\nu,\tau) = \frac{y^{\nu-1}f_T(z)}{\mu^{\nu}\sigma F_T\left[(\sigma|\nu|)^{-1}\right]},$$

em que $y > 0$, $\mu > 0$ é o parâmetro de escala e aproximadamente a mediana; $\sigma > 0$ é apro-

ximadamente o coeficiente de variação; $\nu \in \mathbb{R}$ é um parâmetro de forma, relacionado à assimetria; $\tau > 0$ é um parâmetro de forma, relacionado à curtose; $z = 1/(\sigma\nu)\left[(y/\mu)^{\nu} - 1\right]$, se $\nu \neq 0$, e $z = 1/(\sigma)\log(y/\mu)$, se $\nu = 0$; $f_T(\cdot)$ é a função densidade de probabilidade de uma distribuição PE$(0, 1, \tau)$; $F_T(\cdot)$ é a função de distribuição acumulada de uma distribuição PE$(0, 1, \tau)$. Denota-se $Y \sim$ BCPE(μ, σ, ν) quando a variável aleatória Y segue uma distribuição Box-Cox *power exponential*.

Utilizando A distribuição BCPE, é possível modelar a curtose, simetria, localização e escala da distribuição da variável resposta a partir das covariáveis. Esta distribuição consegue assumir comportamento assimétrico à direita e à esquerda, sendo então apropriada para modelar dados com diferentes comportamentos na simetria com variações nas covariáveis, assim como a distribuição BCCG. Utilizando a distribuição BCPE, também é possível modelar diferentes comportamentos da curtose da variável de interesse com variações nas covariáveis. Nos GAMLSS, os parâmetros μ e ν são modelados pelas covariáveis através da função de ligação identidade e os parâmetros σ e τ são modelados a partir da função de ligação logarítmica. Como a distribuição Box-Cox Cole e Green, esta distribuição apresenta melhores ajustes para valores da variável resposta distantes de zero, e deve ser utilizada em tais casos (Rigby et al., 2019).

2.8 Distribuições Contínuas e Mistas com Suporte no Intervalo Unitário

Distribuições contínuas e mistas com suporte no intervalo unitário são comumente utilizadas para modelar taxas, proporções, porcentagens, frações, ou demais dados limitados ao intervalo unitário. Na Tabela 3 são apresentadas todas distribuições contínuas e mistas que possuem suporte no intervalo unitário presentes no pacote *gamlss* do R, além de suas abreviações e respectivos parâmetros.

Tabela 3: Distribuições contínuas e mistas com suporte no intervalo unitário e suas siglas.

Distribuição	Sigla	Distribuição	Sigla
Beta	BE(μ, σ)	Beta Inflacionada em Zero	BEINF0(μ, σ, ν)
Beta Original	BEo(μ, σ)	Beta Inflacionada em Zero Reparametrizada	BEZI(μ, σ, ν)
Beta Generalizada Tipo 1	GB1(μ, σ, ν, τ)	Beta Inflacionada em Um	BEINF1(μ, σ, ν)
Logito-Normal	LOGITNO(μ, σ)	Beta Inflacionada em Um Reparametrizada	BEOI(μ, σ, ν)
Simplex	SIMPLEX(μ, σ)	Beta Inflacionada em Zero e Um	BEINF(μ, σ, ν, τ)

Todas as distribuições de 2 parâmetros apresentadas na Tabela 3 permitem modelar localização e dispersão. Das distribuições citadas, a distribuição beta é a mais

comumente utilizada para modelagem de dados. Entretanto, algumas adversidades podem surgir quando há necessidade de modelar a assimetria e curtose da variável resposta independentemente da localização e escala. A distribuição Beta Generalizada Tipo 1 permite modelagem da assimetria e curtose de forma independente. Também, as distribuições beta inflacionada em zero, beta inflacionada em um, e beta inflacionada em zero e um permitem o ajuste de modelos para dados no intervalo unitário com excesso de zeros e/ou uns. As distribuições contidas no pacote conseguem assumir forma de "J"ou "U", e adicionalmente a distribuição simplex consegue apresentar bimodalidade. A seguir é detalhada a distribuição beta.

2.8.1 Beta

A distribuição beta é muito utilizada no contexto de modelos de regressão com variável resposta limitada ao intervalo (0,1). A função de densidade de probabilidade em uma parametrização que torna o ajuste sob modelos de regressão interpretável é dada por (Rigby et al., 2019)

$$f(y|\mu,\sigma) = \frac{1}{B(\alpha,\beta)} y^{\alpha-1}(1-y)^{\beta-1},$$

em que $0 < y < 1$, $0 < \mu < 1$ é a média da distribuição; $0 < \sigma < 1$ é um parâmetro de dispersão com efeito multiplicativo no desvio padrão da distribuição; $\alpha = \mu(1-\sigma^2)/\sigma^2$; $\beta = (1-\mu)(1-\sigma^2)/\sigma^2$. Denota-se que uma variável aleatória Y segue uma distribuição Beta por $Y \sim \text{BE}(\mu,\sigma)$.

Na regressão beta e no GAMLSS, modela-se μ e σ a partir da função de ligação logito. Além da inabilidade de modelar a simetria e a curtose independentemente da média e variância, a distribuição beta possui como limitação a impossibilidade de modelar dados com valores iguais a zero e/ou um, necessitando assim a utilização de distribuições inflacionadas com base na distribuição beta original (Rigby et al., 2019).

2.9 Distribuições Discretas de Contagem

Distribuições discretas de contagem são frequentemente empregadas para modelar a ocorrência de eventos, contagens de itens, ou outras quantidades que são limitadas ao conjunto de valores inteiros não negativos. Na Tabela 4 são apresentadas todas as distribuições discretas de contagem presentes no pacote *gamlss* do R, suas abreviações e respectivos parâmetros.

Tabela 4: Distribuições discretas de contagem e suas siglas.

Distribuição	Sigla	Distribuição	Sigla
Beta-Binomial Negativa	BNB(μ, σ, ν)	Yule	YULE(μ)
Delaporte	DEL(μ, σ, ν)	Zipf	ZIPF(μ)
Burr XII Discreta	DBURR12(μ, σ, ν)	Beta-Binomial Negativa Ajustada em Zero	ZABNB(μ, σ, ν, τ)
Poisson Dupla	DPO(μ, σ)	Logarítmica Ajustada em Zero	ZALG(μ, σ)
Poisson Generalizada	GPO(μ, σ)	Binomial Negativa Ajustada em Zero	ZANBI(μ, σ, ν)
Geométrica	GEOM(μ)	Poisson-Gaussiana Inversa Ajustada em Zero	ZAPIG(μ, σ, ν)
Geométrica Original	GEOMo(μ)	Sichel Ajustada em Zero	ZASICHEL(μ, σ, ν, τ)
Logarítmica	LG(μ)	Poisson Ajustada em Zero	ZAP(μ, σ)
Binomial Negativa	NBI(μ, σ)	Zipf Ajustada em Zero	ZAZIPF(μ, σ)
Binomial Negativa Reparametrizada	NBII(μ, σ)	Beta-Binomial Negativa Inflacionada em Zero	ZIBNB(μ, σ, ν, τ)
Família Binomial Negativa	NBF(μ, σ, ν)	Binomial Negativa Inflacionada em Zero	ZINBI(μ, σ, ν)
Poisson	PO(μ)	Família Binomial Negativa Inflacionada em Zero	ZINBF(μ, σ, ν, τ)
Poisson-Gaussiana Inversa	PIG(μ, σ)	Poisson Inflacionada em Zero	ZIP(μ, σ)
Poisson-Gaussiana Inversa Reparametrizada	PIG2(μ, σ)	Poisson Inflacionada em Zero Reparametrizada	ZIP2(μ, σ)
Sichel	SI(μ, σ, ν)	Poisson-Gaussiana Inversa Inflacionada em Zero	ZIPIG(μ, σ, ν)
Sichel Reparametrizada	SICHEL(μ, σ, ν)	Sichel Inflacionada em Zero	ZISICHEL(μ, σ, ν, τ)
Waring	WARING(μ, σ)		

Mais comumente, ajustes para dados dessa natureza são feitos utilizando distribuição Poisson. As maiores dificuldades em modelar dados de contagem ocorrem em casos que a variável resposta possui subdispersão ou sobredispersão em relação à distribuição Poisson, ou quando há excesso ou falta de zeros. Para esse último caso, a implementação de distribuições inflacionadas ou ajustadas em zero conseguem resolver esse problema. Nas seguintes subseções são detalhadas algumas das distribuições apresentadas na Tabela 4 comumente utilizadas para modelagem, bem como aquelas que oferecem flexibilidade para lidar com as adversidades mencionadas.

2.9.1 Poisson

A distribuição Poisson é uma distribuição discreta adequada para dados de contagem. Sua função de probabilidade é dada por (Johnson; Kemp; Kotz, 2005)

$$f(y|\mu) = \frac{\mu^y}{y!}\exp(-\mu),$$

em que $y \in \{0, 1, \ldots\}$, $\mu > 0$ é a média, variância, parâmetro de assimetria e parâmetro de curtose. Quando uma variável aleatória Y segue distribuição Poisson, denota-se $Y \sim \text{PO}(\mu)$. A sua função de probabilidades pode ser reescrita na forma

$$f(y|\mu) = \exp\left(y\log(\mu) - \mu - \log(y!)\right).$$

Assim, a distribuição Poisson pertence à família exponencial uniparamétrica com

- $\theta = \log(\mu)$;
- $b(\theta) = e^{\theta}$, e $\mathrm{E}(Y) = b'(\theta) = \mu$;
- $\phi = 1$, e $\mathrm{Var}(Y) = \phi b''(\theta) = \mu$;
- $c(y, \phi) = \log(y!)$.

Nos MLG e GAMLSS, modela-se μ a partir da função de ligação logarítmica. Como a variância depende apenas da média, essa distribuição deve ser utilizada em dados heterocedásticos. Entretanto, a distribuição Poisson possui várias limitações em suas aplicações. Esse pressupõe equidispersão, ou seja, média igual a variância. Sendo assim, não se ajusta bem quando existe subdispersão ou sobredispersão. Além disso, a distribuição Poisson possui inadequabilidade a dados com alta assimetria, e a impossibilidade de modelar dados com excesso ou escassez de zeros (Rigby et al., 2019).

2.9.2 Binomial Negativa

A distribuição binomial negativa é uma distribuição discreta utilizada para dados de contagem. No contexto de regressão, sua função de densidade de probabilidade pode ser parametrizada como (Johnson; Kemp; Kotz, 2005)

$$f(y|\mu, \sigma) = \frac{\Gamma(y + \frac{1}{\sigma})}{\Gamma(\frac{1}{\sigma})\Gamma(y+1)} \left(\frac{\sigma\mu}{1+\sigma\mu}\right)^{y} \left(\frac{1}{1+\sigma\mu}\right)^{1/\sigma},$$

em que $y \in \{0, 1, \ldots\}$, $\mu > 0$ é a média; $\sigma > 0$ é um parâmetro de dispersão. Quando uma variável aleatória Y segue distribuição binomial negativa, denota-se $Y \sim \mathrm{NBI}(\mu, \sigma)$.

Nos GAMLSS, modela-se μ e σ a partir da função de ligação logarítmica. A distribuição NBI é usualmente utilizada para lidar com sobredispersão em relação a distribuição Poisson, por possuir variância $\mu + \sigma\mu^2$. Assim, o parâmetro σ consegue modelar quanta sobredispersão a distribuição binomial negativa tem em relação à distribuição Poisson, sendo que a distribuição binomial negativa se aproxima da distribuição Poisson quando σ tende a zero (Rigby et al., 2019).

2.9.3 Poisson Dupla

A distribuição Poisson dupla é uma distribuição discreta adequada para dados de contagem. Sua função de densidade de probabilidade é dada por (Rigby et al., 2019)

$$f(y|\mu,\sigma) = c(\mu,\sigma)\sigma^{-1/2}e^{-\mu/\sigma}\left(\frac{\mu}{y}\right)^{y/\sigma}\frac{e^{y/\sigma-y}y^{y}}{y!},$$

em que $y \in \{0,1,\ldots\}$, $\mu > 0$ é aproximadamente a média; $\sigma > 0$ é um parâmetro de dispersão; $c(\mu,\sigma)$ é uma constante de normalização. Quando uma variável aleatória Y segue distribuição Poisson Dupla, denota-se $Y \sim \text{DPO}(\mu,\sigma)$.

Nos GAMLSS, modela-se μ e σ a partir da função de ligação logarítmica. A partir da distribuição Poisson dupla, é possível modelar subdispersão e sobredispersão em relação à distribuição Poisson, uma vez que essa possui média aproximadamente igual a μ, e variância aproximadamente $\sigma\mu$. A possibilidade de modelar subdispersão é a principal vantagem da distribuição DPO em relação à distribuição NBI. O parâmetro σ consegue modelar quanta subdispersão ou sobredispersão a distribuição Poisson dupla tem em relação à distribuição Poisson, sendo que para $\sigma > 1$ a distribuição é aproximadamente uma distribuição Poisson com sobredispersão, para $\sigma < 1$ a distribuição é aproximadamente uma distribuição Poisson com subdispersão, e para $\sigma = 1$ a distribuição é equivalente a uma distribuição Poisson (Rigby et al., 2019).

2.10 Distribuições Discretas de Amplitude Finita

Distribuições discretas limitadas, são comumente utilizadas para modelar contagens de sucessos em um número fixo de tentativas, proporções de ocorrências, ou outras quantidades que são limitadas a um intervalo específico de valores inteiros. A Tabela 5 lista todas as distribuições discretas de amplitude finita disponíveis no pacote *gamlss* do R, incluindo suas abreviações e respectivos parâmetros.

Tabela 5: Distribuições discretas de amplitude finita e suas siglas.

Distribuição	Sigla	Distribuição	Sigla
Binomial	BI(μ)	Binomial Ajustada em Zero	ZABI(μ,σ)
Beta-Binomial	BB(μ,σ)	Beta-Binomial Inflacionada em Zero	ZIBB(μ,σ,ν)
Binomial Dupla	DBI(μ,σ)	Binomial Inflacionada em Zero	ZIBI(μ,σ)
Beta-Binomial Ajustada em Zero	ZABB(μ,σ,ν)		

A distribuição binomial é a mais comumente utilizada para modelar dados discretos com amplitude finita. No entanto, a modelagem de dados dessa natureza pode enfrentar desafios, como excesso ou falta de zeros, subdispersão ou sobredispersão em relação ao que é esperado pela distribuição binomial. Para lidar com sobredispersão em relação à binomial, a distribuição beta-binomial pode ser utilizada, pois esta última in-

corpora variabilidade extra ao considerar a probabilidade de sucesso como uma variável aleatória. Por outro lado, a distribuição binomial dupla é capaz de modelar tanto subdispersão quanto sobredispersão em relação à binomial, oferecendo uma flexibilidade maior para ajustar os dados. Para resolver o problema de excesso ou falta de zeros, podem ser utilizadas distribuições inflacionadas em zero ou ajustadas em zero, acomodando uma quantidade anormal de zeros, e proporcionando uma modelagem mais precisa para conjuntos de dados que apresentam esse tipo de característica.

Além disso, modelos baseados na distribuição binomial também são amplamente utilizados para modelar variáveis dicotômicas. Modelos como modelos de regressão logística e probito são utilizados para modelar a probabilidade de ocorrência de um dos dois possíveis resultados da variável resposta. A distribuição binomial é detalhada a seguir.

2.10.1 Binomial

A distribuição binomial representa o número de sucessos y dado em uma sequência independente de n ensaios de Bernoulli, cada um com probabilidade de sucesso μ, onde n é conhecido. Será utilizada a mesma parametrização apresentada por Rigby et al. (2019), sendo sua função de probabilidade dada por

$$f(y|n,\mu) = \binom{n}{y}\mu^y(1-\mu)^{n-y},$$

em que $y = 0, 1, \ldots, n$, $0 < \mu < 1$ é o parâmetro associado a medidas de tendência central. Quando uma variável aleatória Y segue distribuição binomial, denota-se $Y \sim \text{BI}(n, \mu)$. A sua função de probabilidades pode ser reescrita na forma

$$f(y|n,\mu) = \exp\left\{y\log\left(\frac{\mu}{1-\mu}\right) + n\log(1-\mu) + \log\binom{n}{y}\right\}.$$

A distribuição binomial pertence à família exponencial uniparamétrica com

- $\theta = \log\left(\frac{\mu}{1-\mu}\right) = \text{logit}(\mu)$;
- $b(\theta) = n\log(1+e^\theta)$, e $\text{E}(Y) = b'(\theta) = n\mu$;
- $\phi = 1$, e $\text{Var}(Y) = \phi b''(\theta) = n\mu(1-\mu)$;
- $c(y,\phi) = \log\binom{n}{y}$.

A distribuição binomial é utilizada para modelar variáveis resposta discretas contidas em um intervalo de 0 a n, sendo n conhecido. Um caso particular ocorre quando

$n = 1$, e temos que cada observação da variável resposta possui estado de sucesso ou fracasso. Denomina-se essa distribuição de Bernoulli. Nesse caso, μ representa a probabilidade de sucesso de ocorrência de um evento de interesse, e pelos MLG ou GAMLSS, modela-se μ a partir das covariáveis. Comumente, utiliza-se a função de ligação logito quando modelando μ para MLG e GAMLSS. A distribuição BI não é adequada para modelagem quando a variável resposta dada as covariáveis possui excesso de zeros, ou algum outro comportamento não típico da distribuição binomial (Rigby et al., 2019).

2.11 Modelo de Regressão Linear Normal

O modelo de regressão linear normal é utilizado quando se tem o interesse de estabelecer a relação de uma variável resposta quantitativa Y com um conjunto de variáveis explicativas, também chamadas de covariáveis. O principal interesse do modelo é entender como o conjunto de variáveis explicativas $X_1, X_2, \ldots, X_k$ afetam a média da variável Y a partir de um conjunto de parâmetros desconhecidos $\beta_0, \beta_1, \ldots, \beta_k$, que devem ser estimados a partir de uma amostra (Kutner et al., 2005). Para este modelo, assume-se que as diferentes observações de Y são independentes e seguem uma distribuição normal, com a mesma variância σ^2, e com média dependendo de uma combinação linear das covariáveis com os parâmetros desconhecidos. Assim, a RLN pode ser definida por

$$Y_i \overset{\text{ind}}{\sim} \text{NO}(\mu_i, \sigma^2)$$

$$\mu_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik}$$

em que

- $\overset{\text{ind}}{\sim}$ denota que possuem distribuição de forma independente;
- Y_i é a i-ésima variável resposta;
- μ_i é a média da i-ésima variável resposta;
- $\beta_0, \beta_1, \ldots, \beta_k$ são os $k + 1$ parâmetros (coeficientes) desconhecidos do modelo que relacionam as variáveis explicativas com a média da variável resposta;
- $X_{i1}, X_{i2}, \ldots, X_{ik}$ são os valores fixados das k variáveis explicativas para a i-ésima observação.

Nota-se que a estrutura dos modelos de regressão normal pode ser dividida em dois principais componentes. O primeiro componente se refere a parte aleatória do modelo,

e define a distribuição de Y_i, no caso uma distribuição normal. O segundo componente se refere à parte sistemática (determinística) do modelo, e define a relação que as variáveis explicativas possuem com a média da variável resposta, no caso, uma relação linear com os parâmetros regressores $\beta_0, \beta_1, \ldots, \beta_k$.

O modelo de RLN pode ainda ser representado de forma matricial dada por

$$\mathbf{Y} \overset{\text{ind}}{\sim} \text{NO}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$$

$$\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$$

em que

- $\mathbf{Y}$ é o vetor de dimensão n contendo as n variáveis resposta;
- a notação $\text{NO}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$ indica que cada variável Y_i do vetor $\mathbf{Y}$ segue uma distribuição normal com média μ_i do vetor $\boldsymbol{\mu}$ e variância σ_i^2 do vetor $\boldsymbol{\sigma}^2$, onde $\sigma_i^2 = \sigma^2$, para todo $i \in \{1, 2, \ldots, n\}$;
- $\boldsymbol{\beta}$ é o vetor dos $k+1$ coeficientes desconhecidos do modelo;
- $\mathbf{X}$ é uma matriz de dimensão n por $k+1$ de valores conhecidos, em que a primeira coluna é composta por 1's e as demais são compostas pelos valores fixados das k variáveis explicativas.

Essa será a notação matricial utilizada nos demais casos afrente.

Junto com sua definição, o modelo de RLN apresenta algumas suposições que podem limitar ou dificultar seu ajuste aos dados. Primeiro, o modelo assume que a variável resposta segue uma distribuição normal, o que pode ser uma suposição equivocada dependendo da natureza da variável resposta, como no caso de variáveis definidas em um domínio limitado, ou variáveis com uma natureza assimétrica. Segundo, assume-se também que a variância da variável resposta é constante independentemente dos valores das variáveis explicativas, o que também pode ser uma suposição inadequada. O modelo de RLN também supõe que a relação entre a média da variável resposta e as variáveis explicativas possuem uma relação linear, o que pode gerar conclusões errôneas no caso de uma relação mais complexa ou quando a média deve ser limitada a um certo domínio de valores. Os modelos discutidos a seguir visam solucionar alguns desses problemas.

2.12 Modelo Linear Generalizado

Os modelos lineares generalizados, similarmente aos modelos de RLN, visam entender o comportamento de uma variável resposta Y a partir de covariáveis. Entretanto, nos MLGs, a variável resposta pode assumir um leque maior de distribuições podendo pertencer a qualquer distribuição da família exponencial uniparamétrica (Paula, 2023). A partir da definição da família exponencial uniparamétrica dada na Seção 2.1, pode-se definir a estrutura dos MLGs a partir de

$$\mathbf{Y} \overset{\text{ind}}{\sim} \text{FE}(\boldsymbol{\mu}, \boldsymbol{\phi})$$

$$\boldsymbol{\mu} = g^{-1}(\boldsymbol{\eta})$$

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$$

em que

- $\mathbf{Y}$ é o vetor de variáveis resposta de dimensão n por 1, com média $\boldsymbol{\mu}$ e parâmetro de dispersão $\boldsymbol{\phi}$;
- FE representa alguma distribuição da família exponencial uniparamétrica, com média μ e parâmetro de dispersão ϕ;
- $\boldsymbol{\eta}$ é o preditor linear de dimensão n por 1;
- $g(\cdot)$ é a função de ligação que associa o preditor linear com a média da variável resposta;
- $\boldsymbol{\beta}$ é o vetor dos $k+1$ coeficientes desconhecidos do modelo;
- $\mathbf{X}$ é uma matriz de dimensão n por $k+1$ de valores conhecidos, em que a primeira coluna é composta por 1's e as demais são compostas pelos valores fixados das k variáveis explicativas.

Nota-se que a estrutura dos MLGs pode ser dividida em três principais componentes. O primeiro se refere a parte aleatória do modelo, e define a distribuição de Y_i, podendo ser qualquer distribuição da família exponencial uniparamétrica. O segundo componente se refere à parte sistemática do modelo, também chamado de preditor, e é definido pela multiplicação dos parâmetros de regressão e as variáveis explicativas. Por fim, há um novo componente no modelo chamado função de ligação. Este define a relação

existente entre o preditor e a média da variável resposta. Consequentemente, pode-se acomodar uma natureza mais diversificada de variáveis resposta, atendendo limitações dos valores que estas podem assumir. Também, a partir da função de ligação, respeita-se o domínio dos parâmetros da distribuição sem precisar limitar os parâmetros de regressão, possibilitando uma maior liberdade no ajuste dos modelos. Com isso, os MLGs possuem uma maior capacidade de ajuste comparado aos modelos de RLN.

Mesmo com a versatilidade dos MLGs, estes ainda possuem algumas limitações. A primeira surge da necessidade da variável resposta ser parte da família exponencial uniparamétrica, excluindo assim a sua aplicação para variáveis resposta que podem seguir distribuições fora desta família. A segunda surge a partir da limitação da modelagem para um único parâmetro da distribuição, o que limita os modelos a classe da família exponencial uniparamétrica. Por fim, tem-se a limitação da parte sistemática ser linear nos parâmetros regressores, o que impede relações mais sofisticadas entre as variáveis explicativas e o preditor.

2.13 Modelos Aditivos Generalizados de Localização, Escala e Forma

Rigby e Stasinopoulos (2005) propuseram a classe GAMLSS que pode acomodar possíveis limitações dos MLGs, além de incorporarem diversas outras classes de modelos. A partir dos GAMLSS, tem-se escolhas muito mais abrangente de distribuições de probabilidade para a variável resposta. De fato, nos GAMLSS, a resposta Y pode assumir qualquer distribuição de até 4 parâmetros, contanto que a função densidade de probabilidade seja diferenciável em termos dos parâmetros, ou analiticamente ou numericamente (Stasinopoulos et al., 2017). Além disso, os GAMLSS permitem que todos os 4 parâmetros da distribuição sejam modelados a partir de covariáveis, permitindo assim um nível de explicação muito maior da relação entre a variável resposta e as variáveis explicativas. Por fim, essa classe de modelos permite que o preditor possua não apenas termos lineares, como também termos não lineares, termos não paramétricos e até mesmo termos de efeitos aleatórios, possibilitando uma relação muito mais flexível entre as covariáveis e o preditor. Sendo assim, será possível escrever modelos mistos ou modelos aditivos generalizados como casos específicos dos GAMLSS. Apesar disto, nesse trabalho haverá um foco maior nos casos em que o preditor é linear, paramétrico e sem efeitos aleatórios.

Os GAMLSS podem ser definido por

$$\mathbf{Y} \overset{\text{ind}}{\sim} \mathrm{D}(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\nu}, \boldsymbol{\tau})$$

$$\boldsymbol{\eta}_1 = g_1(\boldsymbol{\mu}) = \mathbf{X}_1\boldsymbol{\beta}_1 + s_{11}(\mathbf{x}_{11}) + \cdots + s_{1J_1}(\mathbf{x}_{1J_1})$$

$$\boldsymbol{\eta}_2 = g_2(\boldsymbol{\sigma}) = \mathbf{X}_2\boldsymbol{\beta}_2 + s_{21}(\mathbf{x}_{21}) + \cdots + s_{2J_2}(\mathbf{x}_{2J_2})$$

$$\boldsymbol{\eta}_3 = g_3(\boldsymbol{\nu}) = \mathbf{X}_3\boldsymbol{\beta}_3 + s_{31}(\mathbf{x}_{31}) + \cdots + s_{3J_3}(\mathbf{x}_{3J_3})$$

$$\boldsymbol{\eta}_4 = g_4(\boldsymbol{\tau}) = \mathbf{X}_4\boldsymbol{\beta}_4 + s_{41}(\mathbf{x}_{41}) + \cdots + s_{4J_4}(\mathbf{x}_{4J_4})$$

em que

- $\mathbf{Y}$ é o vetor de variáveis resposta de dimensão n por 1;
- $\boldsymbol{\mu}$ é o vetor de parâmetros de localização de dimensão n por 1;
- $\boldsymbol{\sigma}$ é o vetor de parâmetros de escala de dimensão n por 1;
- $\boldsymbol{\nu}$ é o vetor de parâmetros associado ao primeiro parâmetro de forma de dimensão n por 1;
- $\boldsymbol{\tau}$ é o vetor de parâmetros associado ao segundo parâmetro de forma de dimensão n por 1;
- D representa alguma distribuição com no máximo 4 parâmetros diferenciável com relação aos parâmetros;
- $\boldsymbol{\eta}_p$ representa o vetor do preditor de dimensão n por 1 associado ao p-ésimo parâmetro, com $p \in \{1, 2, 3, 4\}$;
- $g_p(\cdot)$ é a função de ligação associada ao preditor do p-ésimo parâmetro;
- $\boldsymbol{\beta}_p$ é o vetor dos parâmetros regressores desconhecidos de dimensão $k_p + 1$ por 1 associado ao p-ésimo parâmetro da distribuição;
- $\mathbf{X}_p$ é uma matriz de valores conhecidos de dimensão n por $k_p + 1$ associado ao p-ésimo parâmetro, em que a primeira coluna é composta por 1's e as demais são compostas pelos valores fixados das respectivas variáveis explicativas;
- $s_{pj}(\mathbf{x}_{pj})$ podem ser termos não lineares, termos não paramétricos ou termos de efeitos aleatórios. Considerando-se $s_{pj}(\mathbf{x}_{pj}) = \mathbf{0}$ para todo p e j, o modelo terá apenas preditores lineares.

Apesar dos GAMLSS comportarem distribuições de até 4 parâmetros, essa classe pode facilmente ser expandida para distribuições com 5 ou mais parâmetros. Também, distribuições além das mencionadas nas Seções 2.6 a 2.10 podem ser incorporadas a partir de transformações de demais distribuições, contanto que as condições descritas nessa seção forem atendidas. Mesmo com a grande flexibilidade proporcionada pelos GAMLSS, estes ainda possuem suposições em suas aplicações que devem ser analisadas, uma vez que quebras das suposições podem levar a estimativas viesadas e inferências incorretas.

Uma dessas é a suposição de independência das observações. Essa suposição nem sempre é válida, especialmente em dados temporais ou espaciais, onde pode haver correlação entre observações. Para contornar essa limitação, pode ser necessário ajustar modelos que levem em conta a correlação, como com a inclusão de efeitos aleatórios ou estruturas de series temporais, embora isso possa complicar consideravelmente o modelo e suas interpretações.

A adequação do modelo indicado aos dados também é uma suposição dos GAMLSS que deve ser cumprida. Se a distribuição escolhida não for adequada para os dados, ou se as funções preditoras não capturarem adequadamente as relações entre as variáveis, o modelo pode não fornecer boas estimativas ou previsões. A flexibilidade dos GAMLSS, embora poderosa, também aumenta a complexidade da modelagem, exigindo um conhecimento profundo das características importantes dos dados, e se essas características estão sendo devidamente modeladas.

A robustez dos modelos à *outliers* ou pontos extremos também é um aspecto necessário de se analisar ao utilizar GAMLSS, especialmente quando a distribuição de resposta não é robusta a valores extremos. *Outliers* podem influenciar fortemente as estimativas dos parâmetros, levando a conclusões incorretas. Assim, é importante realizar análises de diagnóstico para identificar *outliers*, e considerar técnicas robustas ou transformações que possam mitigar o impacto desses. Rigby et al. (2019) e Aeberhard et al. (2021) investigam estimações robustas dos GAMLSS, havendo no pacote *GJRM* do R (Marra; Radice, 2023) implementação de ajustes robustos para essa classe de modelos.

Ainda, existem extensões dos GAMLSS para resolver problemas específicos, e pacotes no R que auxiliam na resolução de problemas mais complexos. Umlauf, Klein e Zeileis (2018) apresentam ajustes bayesianos para os GAMLSS, com implementação disponível a partir do pacote *bamlss* no R. Também, o pacote *GJRM* permite o uso de copulas para implementação de variáveis resposta multivariadas nos GAMLSS. O pacote *VGAM* (Yee, 2024) apresenta utilização de modelos aditivos generalizados vetoriais. Hofner, Mayr e Schmid (2016) apresentam técnicas de seleção de variáveis e construção de

modelos para GAMLSS a partir da técnica de *boosting*, pelo pacote *gamboostLSS* no R.

2.14 Estimação e Inferência para os GAMLSS

Para os GAMLSS, a estimação dos parâmetros é feita pelo método de máxima verossimilhança. O logaritmo da função de verossimilhança sob os GAMLSS paramétricos, dada a independência das observações, é dada por

$$\ell = \sum_{i=1}^{n} \log f(y_i|\mu_i, \sigma_i, \nu_i, \tau_i),$$

em que $f(y_i|\mu_i, \sigma_i, \nu_i, \tau_i)$ é a função densidade de probabilidade da i-ésima observação da variável resposta com parâmetros $\mu_i, \sigma_i, \nu_i, \tau_i$.

Existem dois algoritmos para a maximização do logaritmo da função de verossimilhança utilizados para os GAMLSS, o algoritmo Rigby e Stasinopoulos (RS) e o algoritmo Cole e Green (CG). Ambos algoritmos consistem em chamar repetidamente um algoritmo de mínimos quadrados ponderados iterativos, até que a convergência do desvio global seja alcançada. Rigby e Stasinopoulos (2005) apresentam tais algoritmos detalhadamente, e provam que ambos convergem com estimadores de máxima verossimilhança.

Stasinopoulos et al. (2017) comentam que o algoritmo CG é em ocasiões instável, especialmente nas primeiras iterações do algoritmo, o que pode levar a divergência das estimativas. Também, os autores comentam que o algoritmo RS é no geral muito mais estável e em muitos casos mais rápido, mas que em casos dos parâmetros ajustados da distribuições serem altamente correlacionados, o algoritmo RS pode ser mais devagar, e convergir antes de chegar nas estimativas de máxima verossimilhança. Nesses casos, pode ser aumentado o número de iterações do algoritmo, ou até utilizar uma mistura dos algoritmos CG e RS.

Uma vez obtidas as estimativas pontuais de máxima verossimilhança dos coeficientes do modelo, torna-se necessário o entendimento da variabilidade desses estimadores para se avaliar a significância estatística de cada coeficiente do modelo e construir intervalos de confiança. Uma vez que foram obtidos estimadores de máxima verossimilhança, sob modelo adequado, espera-se que os coeficientes sejam consistentes, invariantes a parametrizações, assintoticamente eficientes e assintoticamente normais. Seja $\boldsymbol{\theta}$ o conjunto de todos os coeficientes lineares para os parâmetros $\mu_i, \sigma_i, \nu_i, \tau_i$, ou seja, $\boldsymbol{\theta} = (\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top, \boldsymbol{\beta}_3^\top, \boldsymbol{\beta}_4^\top)^\top$. Segue que, assintoticamente,

$$\hat{\boldsymbol{\theta}} \sim \text{NO}(\boldsymbol{\theta}_T, \mathbf{i}(\boldsymbol{\theta}_T)^{-1}),$$

em que $\hat{\boldsymbol{\theta}}$ é o estimador de máxima verossimilhança de $\boldsymbol{\theta}$, $\boldsymbol{\theta}_T$ é o valor real dos parâmetros, e

$$\mathbf{i}(\boldsymbol{\theta}_T) = -\text{E}\left[\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top}\right]_{\boldsymbol{\theta}_T}$$

é a matriz de informação esperada de Fisher avaliada no valor real dos parâmetros $\boldsymbol{\theta}_T$. Uma vez que nem sempre é possível calcular essa matriz analiticamente, a matriz de informação observada $\mathbf{I}(\boldsymbol{\theta}_T)$ definida como

$$\mathbf{I}(\boldsymbol{\theta}_T) = -\left[\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top}\right]_{\boldsymbol{\theta}_T}$$

é utilizada em seu lugar. Uma vez que $\boldsymbol{\theta}_T$ é desconhecido, substitui-se $\boldsymbol{\theta}_T$ por $\hat{\boldsymbol{\theta}}$ tanto para a matriz de informação observada quanto para a matriz de informação esperada, obtendo-se $\mathbf{I}(\hat{\boldsymbol{\theta}})$ e $\mathbf{i}(\hat{\boldsymbol{\theta}})$. Usualmente nos GAMLSS paramétricos, utiliza-se a seguinte distribuição assintótica para $\hat{\boldsymbol{\theta}}$:

$$\hat{\boldsymbol{\theta}} \sim \text{NO}(\boldsymbol{\theta}_T, \mathbf{I}(\hat{\boldsymbol{\theta}})^{-1}).$$

A partir desse resultado, será possível fazer predições intervalares utilizando o modelo, construir intervalos de confiança Wald para os coeficientes e testar significância de variáveis a partir de testes Wald. Stasinopoulos et al. (2017) sugerem ainda que é possível obter estimativas do erro padrão dos coeficientes por *bootstrap*. Adicionalmente, a fim de se ter mais robustez em predições, é possível ajustar vários modelos ao mesmo banco e obter predições a partir da média dos valores preditos de cada modelo.

2.15 Critérios de Seleção de Modelos para os GAMLSS

A escolha de um modelo de regressão consiste na escolha de uma distribuição adequada, escolha das variáveis dependentes para cada parâmetro da distribuição da variável resposta, e escolha de como os termos afetam os parâmetros, tanto pela escolha adequada de uma função de ligação quanto na adição de transformações das variáveis explicativas. Tais escolhas podem afetar diretamente na performance dos modelos, sendo que a não especificação correta do modelo pode levar a casos de *overfit* (quando se interpreta exces-

sivamente os dados) ou *underfit* (quando se interpreta escassamente os dados), trazendo problemas para predição ou interpretação. Assim, se torna crucial a seleção adequada dos modelos. Entretanto, não existe método automático para escolha de distribuição e termos (Stasinopoulos et al., 2024) nos GAMLSS, sendo necessário encontrar o devido balanço entre o viés e variância dos modelos ajustados. Normalmente, modelos mais complexos apresentam maior variância porém menor viés, e modelos menos complexos apresentam menor variância, porém maior viés. Algumas técnicas e critérios diferentes podem ser utilizadas para comparação de modelos.

O desvio global $\text{GDEV} = -2\ell(\hat{\boldsymbol{\theta}})$ pode ser utilizado para comparação entre modelos encaixados, ou seja, quando um modelo é caso especial de outro (Stasinopoulos et al., 2017). Assim, sendo GDEV_1 o desvio global de um modelo mais complexo, e GDEV_0 o desvio global de um modelo menos complexo, $\Lambda = \text{GDEV}_0 - \text{GDEV}_1$ segue distribuição qui-quadrado com $d = df_0 - df_1$ graus de liberdade, sob hipótese nula de que o modelo mais simples é o modelo adequado. Assim, pode ser construído um teste de hipótese para comparar os dois modelos.

Ainda, Akaike (1983) propôs o critério de Akaike generalizado (GAIC) para comparação de modelos penalizando *overfit*, obtido a partir de $\text{GAIC}(\kappa) = \text{GDEV} + (\kappa \times df)$ onde df denota o número de graus de liberdade usados no modelo, e κ é um termo de penalização para a complexidade do modelo. Escolhe-se o modelo com menor valor para o GAIC. O critério de informação de Akaike (AIC) (Akaike, 1974) e o critério de informação bayesiano (BIC) (Schwarz, 1978) são casos particulares do GAIC, para $\kappa = 2$ e $\kappa = \log(n)$, respectivamente. Stasinopoulos et al. (2017) sugerem que o AIC tende a cometer *overfit* na seleção de modelos, e o BIC tende a cometer *underfit* na seleção de modelo, assim sendo recomendado o uso de um valor de $2,5 \leq \kappa \leq 4$. Esse critério equivale a uma penalização da função de verossimilhança conforme o modelo se torne mais complexo. Apesar desse critério ser apropriado para comparação de modelos encaixados, e modelos não encaixados possuírem funções de verossimilhança diferentes, Stasinopoulos et al. (2017) ainda sugerem a utilização do critério para modelos não encaixados. No atual trabalho, o critério será empregado para obter uma indicação sobre quais distribuições são mais adequadas, contudo ele não será decisivo para a escolha entre modelos não encaixados, sendo então utilizadas outras técnicas para esses casos.

Com o intuito de fornecer uma indicação das possíveis características da variável resposta e das distribuições que podem ser adequadas para essas características, a função *fitDist()* do pacote *gamlss* ajusta marginalmente todas as distribuições no suporte da variável resposta e retorna o GAIC para cada distribuição. Entretanto, essa função é uti-

lizada apenas com intuito sugestivo, uma vez que o comportamento marginal da variável resposta pode diferir significativamente do comportamento condicional às covariáveis do modelo.

Ainda, podem ser implementadas técnicas de validação cruzada ou separação do banco em treino, validação e teste para a escolha de modelos, sendo escolhidos os modelos com melhor métrica de qualidade de ajustamento. Métricas para avaliar a qualidade de ajuste incluem o desvio global ou métricas de dispersão.

No contexto dos GAMLSS, a tarefa de escolha de modelo pode ser subdividida em 4 escolhas diferentes: a escolha da distribuição, a escolha das funções de ligação para cada parâmetro da distribuição da variável resposta, a escolha dos termos preditores para cada preditor, e valores de hiperparâmetros. Uma vez que o atual trabalho apenas analisa os GAMLSS no contexto paramétrico, não é necessária a etapa de seleção de hiperparâmetros.

Para a escolha da distribuição adequada, Rigby et al. (2019) destacam como importante serem considerados o suporte da distribuição; a interpretação do parâmetro de localização (média, mediana, moda, etc.); a existência de medidas de momento (média, variância, assimetria, curtose) em termos dos parâmetros da distribuição, que auxiliam na interpretabilidade do modelo; flexibilidade em modelar assimetria e curtose, especialmente quando o interesse da modelagem está nas caudas da variável resposta; robustez dos estimadores de máxima verossimilhança para *outliers*; a existência de medidas explícitas de centis e medidas baseadas em centis de localização, escala, assimetria e curtose, quando o foco está nos quantis da distribuição. A escolha de uma distribuição adequada pode ser validada através de um *worm plot* (discutido na próxima subseção).

Ao escolher uma função de ligação adequada para o modelo, é necessário avaliar se a função de ligação considerada é adequada ao domínio dos parâmetros, e qual a natureza esperada das variáveis explicativas no parâmetro da distribuição (aditivo, multiplicativo). A adequabilidade da função de ligação pode ser validada pelo *worm plot* (discutido na próxima subseção).

Finalmente, para realizar a seleção de termos ou variáveis nos modelos, existem técnicas baseadas em critérios como o GAIC. A função *stepGAIC()* do pacote *gamlss* performa uma seleção *stepwise* das variáveis para um parâmetro da distribuição da variável resposta, mantendo as demais estruturas de regressão para os demais parâmetros da distribuição. A função *stepGAICAll.A()* performa uma seleção *forward* para cada parâmetro da distribuição da variável resposta, e depois uma seleção *backward* para cada parâmetro

da distribuição da variável resposta, possibilitando diferentes estruturas de regressão para diferentes parâmetros da distribuição. Ainda, tem-se a função *stepGAICAll.B()* que performa uma seleção *stepwise* mantendo as mesmas variáveis em todas as estruturas de regressão dos parâmetros.

2.16 Técnicas de diagnóstico para os GAMLSS

Após o ajuste de um modelo, se torna necessário realizar um diagnóstico para verificar se as suposições feitas são válidas. Sem a devida validação do modelo, as inferências e interpretações fornecidas por este não são confiáveis. Técnicas de diagnóstico também auxiliam na identificação de pontos discrepantes e avaliam a qualidade do ajuste obtido para os dados, garantindo a adequabilidade dos resultados.

A análise de resíduos é comumente utilizada para realizar diagnóstico, pois os resíduos indicam a discrepância entre o modelo ajustado e os dados observados. Na RLN e nos MLGs, são frequentemente utilizados os resíduos estudentizados, resíduos de Pearson, e o resíduo deviance no diagnóstico. No entanto, Stasinopoulos et al. (2017) observam que os resíduos de Pearson podem não seguir um comportamento normal quando a variável resposta é altamente assimétrica ou possui alta curtose. Além disso, esses resíduos não são bem definidos ao modelar múltiplos parâmetros da variável resposta. Dunn e Smyth (1996) propuseram o resíduo quantílico, que é mais adequado no contexto dos GAMLSS. Nesta seção, será introduzido o resíduo quantílico, além de outras ferramentas de diagnóstico baseadas nesse resíduo.

2.16.1 Resíduo Quantílico

Para o diagnóstico dos GAMLSS, o resíduo quantílico é o mais comumente utilizado. Seja o resíduo quantílico r definido como

$$r = \Phi^{-1}(u)$$

em que $\Phi^{-1}(\cdot)$ é a inversa da função de distribuição acumulada de uma distribuição $\text{NO}(0, 1)$ e u é definido de maneira diferente para variáveis resposta contínuas e discretas.

Se a variável resposta Y for contínua, $u = F(y|\boldsymbol{\theta})$, a função de distribuição acumulada do modelo, possui distribuição uniforme contínua no intervalo de zero e um. Assim, o resíduo quantílico r possui distribuição $\text{NO}(0, 1)$ se o modelo for correto para a

variável resposta.

Se a variável resposta Y for discreta, u é definido como um valor aleatório de uma distribuição uniforme no intervalo $[u_1, u_2] = [F(y-1|\boldsymbol{\theta}), F(y|\boldsymbol{\theta})]$. Como no caso anterior, o resíduo quantílico r possui distribuição NO$(0, 1)$ se o modelo for correto para a variável resposta.

Uma vez que os parâmetros da distribuição da variável resposta são desconhecidos, os resíduos quantílicos são definidos como

$$\hat{r}_i = \Phi^{-1}(\hat{u}_i)$$

onde $\hat{u}_i$ é definido de maneira análoga ao descrito acima, onde $\hat{u} = F(y|\hat{\boldsymbol{\theta}})$ caso a variável resposta seja contínua, e $\hat{u}$ é definido como um valor aleatório de uma distribuição uniforme no intervalo $[\hat{u}_1, \hat{u}_2] = [F(y-1|\hat{\boldsymbol{\theta}}), F(y|\hat{\boldsymbol{\theta}})]$, sendo $\hat{\boldsymbol{\theta}}$ o vetor de estimadores de máxima verossimilhança de $\boldsymbol{\theta}$. Caso o modelo especificado seja correto para a variável resposta, $\hat{r}_i$ terá distribuição assintoticamente NO$(0, 1)$ e serão independentes assintoticamente.

A principal vantagem do resíduo quantílico é que independente da distribuição original da variável resposta, se o modelo assumido for o correto, o resíduo quantílico seguirá uma distribuição assintoticamente normal. Isso facilita a implementação de técnicas diagnósticas baseadas na normalidade dos resíduos. Vale notar que no uso de resíduos quantílicos para variáveis discretas, recomenda-se a réplica de vários conjuntos de resíduos quantílicos antes de se chegar a uma conclusão sobre o ajuste do modelo.

Técnicas usuais de análise de resíduos podem ser conduzidas a partir do resíduo quantílico. A função *plot()* do pacote *gamlss* gera gráficos comuns de análise de resíduos para verificar a adequabilidade dos modelos. São apresentados pela função o gráfico de dispersão dos resíduos contra os valores ajustados de μ e o gráfico de dispersão dos resíduos contra os índices, nos quais se espera um comportamento aleatório dos resíduos em torno de zero, demonstrando a independência dos resíduos em relação aos valores ajustados de μ e às observações da variável resposta, além da variância constante dos resíduos. Também são esboçados a estimativa da densidade dos resíduos por kernel e o *QQ-plot* desses, a fim de verificar a normalidade dos resíduos e identificar possíveis observações atípicas.

2.16.2 *Worm plot*

Proposto por Buuren e Fredriks (2001), o *worm plot* é uma ferramenta gráfica utilizada para diagnosticar a adequação de modelos. Este permite a identificação de

padrões nos resíduos que podem indicar desvios de suposições, e a avaliação de se os resíduos quantílicos seguem a distribuição esperada. Além disso, a partir do *worm plot* é possível identificar áreas onde o ajuste do modelo não foi adequado.

O *worm plot* consiste em um *QQ-plot* sem tendência, ou vários *QQ-plot* sem tendência avaliados para diferentes intervalos das covariáveis. Assim, no eixo das abscissas estão os valores esperados dos resíduos quantílicos sob adequação do modelo postulado aos dados, enquanto no eixo das ordenadas estão os valores dos resíduos quantílicos menos o quantil esperado da NO$(0, 1)$ para aquela observação. No gráfico, são apresentados pontos que indicam o quão afastados os valores observados dos resíduos ordenados estão de seus valores teóricos aproximados para cada observação, bandas de 95% de confiança para cada ponto, e um polinômio de terceiro grau ajustado aos pontos do *worm plot*. A Figura 1 apresenta um exemplo de um *worm plot* para um ajuste adequado dos dados.

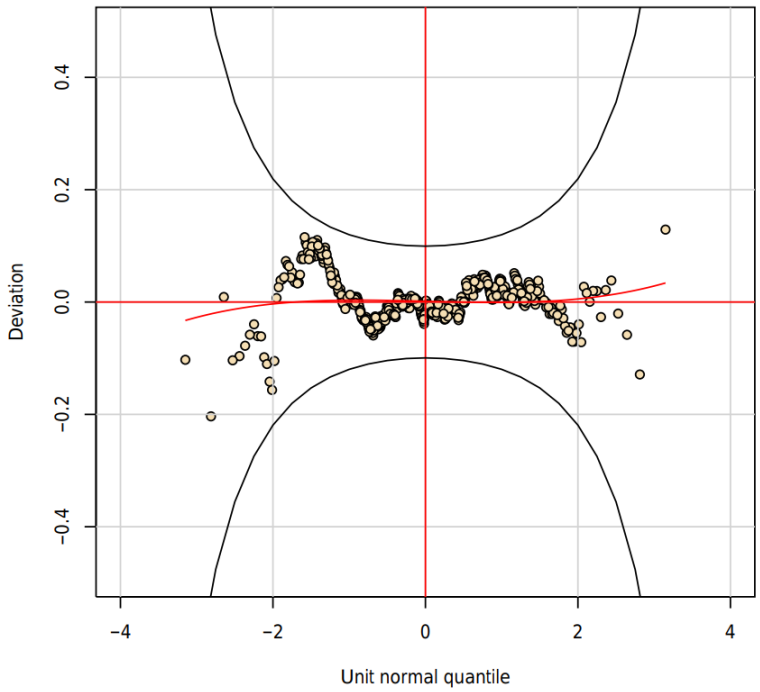

Figura 1: Exemplo de um *worm plot* (Stasinopoulos et al., 2017).

Comportamentos como tendências nos pontos e mais de 5% dos pontos estarem fora das bandas de confiança, indicam falta de ajuste do modelo postulado aos dados. Se houver pontos muito distantes das bandas, considera-se como indícios desses pontos serem *outliers* que prejudicam o ajuste global. Por não haver nenhum desses comportamentos no *worm plot* da Figura 1, conclui-se que a distribuição postulada se ajustou bem aos dados.

Tendências específicas nos pontos e na linha ajustada no *worm plot* podem ser

indícios de certas deficiências nos resíduos, que refletem em deficiências pontuais do modelo ajustado. Estes padrões podem dar ideias sobre o real comportamento da variável resposta, e como melhorar o ajuste aos dados. Assim, Buuren e Fredriks (2001) e Stasinopoulos et al. (2017) citam as seguintes interpretações sobre tendências nos *worm plots* apresentadas na Tabela 6, e suas reflexões sobre a média, variância, assimetria e curtose da distribuição ajustada.[2]

Tabela 6: Interpretação do ajuste de um modelo baseado em diferentes formas do *worm plot*.

Forma do *Worm plot*	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 Sobrestimado
Inclinação Positiva	Variância Muito Alta	Parâmetro de Escala Muito Baixo
Inclinação Negativa	Variância Muito Baixa	Parâmetro de Escala Muito Alto
Forma de U	Assimetria à Direita	Excesso de Assimetria à Esquerda
Forma de U Invertido	Assimetria à Esquerda	Excesso de Assimetria à Direita
Forma de S com Esquerda para Baixo	Leptocúrticos	Caudas Muito Leves
Forma de S com Esquerda para Cima	Platicúrticos	Caudas Muito Pesadas

A partir da interpretação de cada *worm plot*, caso haja alguma tendência na forma, é necessário que um novo modelo seja ajustado com mais flexibilidade nos parâmetros relacionados a cada déficit no modelo original, seja por tornar a função para aquele parâmetro mais flexível ou por utilizar uma distribuição mais flexível. Por exemplo, caso seja ajustado um modelo utilizando uma distribuição simétrica, e o *worm plot* indicar uma assimetria nos resíduos, pode ser necessário o ajuste de um modelo por uma distribuição que acomode assimetria. Em um mesmo ajuste, pode-se ter mais de um comportamento sistemático dos resíduos presente no *worm plot*, embora em tais casos seja difícil identificar mais de uma tendência.

No geral, quanto mais perto os pontos e a linha de tendência do gráfico estiver de uma linha horizontal cruzando a origem, mais próxima é a distribuição dos resíduos de uma distribuição normal padrão, indicando um melhor ajuste do modelo (Stasinopoulos et al., 2017). Assim, esses gráficos podem ser utilizados para comparar diferentes modelos, indicando qual fornece o melhor ajuste quando as curvas estão próximas da linha zero (Buuren, 2007).

[2]Exemplos visuais de cada forma podem ser encontradas a partir da referência Stasinopoulos et al. (2017).

3 Metodologia

O trabalho foi desenvolvido a partir do estudo dos GAMLSS, seguindo as referências Rigby e Stasinopoulos (2005), Stasinopoulos et al. (2017), Rigby et al. (2019) e Stasinopoulos et al. (2024). Inicialmente foi feita uma revisão teórica do modelo de regressão normal, e do modelo linear generalizado, em que compreendeu-se suas utilidades e limitações. Adicionalmente, foi feito um estudo sobre o modelo aditivo generalizado, entretanto esse não foi foco desse trabalho. Então, estudou-se os GAMLSS, e como esse é utilizado para suprir as limitações dos modelos previamente mencionados. Foram estudadas aplicações para variáveis respostas contidas na reta real, e casos da variável resposta ser contínua positiva, limitada a intervalos, discreta e binária. Posteriormente, foi estudada a parte inferencial dos GAMLSS, além de métodos diagnósticos e técnicas para seleção de modelos. Por fim, os modelos estudados foram aplicados a dados reais, a fim de mostrar a versatilidade dos modelos além de comprovar a sua aplicabilidade a dados reais, e estudar possíveis limitações dos modelos.

3.1 Apoio computacional

Todo o eventual processamento de dados, geração de imagens, cálculos, e outras eventuais técnicas computacionais foram feitos com o apoio do software R, a partir da plataforma RStudio (disponível em https://posit.co/products/open-source/rstudio).

3.1.1 Pacote *gamlss*

Toda a aplicação computacional referente aos GAMLSS será feita por meio do pacote *gamlss*, publicado por Stasinopoulos et al. (2023) pelo software R.

4 Aplicações

Feita a revisão teórica, os modelos estudados foram aplicados a dados reais, com o objetivo de destacar as características e propriedades típicas dos GAMLSS descritas no estudo teórico. Assim, os exemplos servirão para destacar vantagens e limitações dos GAMLSS. Os dados utilizados para tais análises foram obtidos de referências voltadas à modelagem por regressão, entretanto foram analisados utilizando a base teórica apresentada neste trabalho.

4.1 Aplicação 1: Preço líquido do aluguel pela área de imóveis

A modelagem por regressão do preço líquido do aluguel de imóveis pode ajudar a identificar possíveis fatores que influenciam no aumento desses preços, ajudando a prever tendências de mercado. O banco *rent* disponível no pacote *gamlss.data* do R é baseado em uma pesquisa realizada em abril de 1993 pela Infratest Sozialforschung, onde foi selecionada uma amostra aleatória de 1967 observações de acomodações em Munique, com novos contratos de locação ou aumentos de aluguel nos últimos quatro anos. Fahrmeir, Gieger e Klinger (1995) e Stasinopoulos, Rigby e Fahrmeir (2000) fizeram análises relacionadas a esse banco, entretanto suas modelagens foram feitas a partir de modelos aditivos, os quais possuem menos interpretabilidade de que modelos totalmente paramétricos. Assim, essa aplicação visa ajustar um modelo para tais dados de forma paramétrica, mantendo a interpretabilidade dos coeficientes.

O banco *rent* possui informações sobre o valor do aluguel mensal em marcos alemães, o tamanho da acomodação em metros quadrados, e demais informações sobre as acomodações. Por fins didáticos, serão apenas consideradas no ajuste dos modelos as variáveis valor do aluguel e o tamanho da acomodação, variável resposta e variável explicativa respectivamente, ainda que a exclusão das demais variáveis possa afetar a qualidade da análise.

A Figura 2 apresenta o gráfico de dispersão entre a variável aluguel e a variável tamanho. A partir do gráfico, percebe-se uma tendência de crescimento do preço do aluguel conforme o tamanho das acomodações aumentam, um comportamento esperado para esses dados. Percebe-se também que a maioria dos dados está concentrada em valores menores de preço e tamanho, havendo algumas poucas observações com valores mais elevados de preço e também de tamanho. Como esperado das variáveis por suas naturezas, ambas apresentam apenas valores positivos, o que pode ser um indício de que

seja necessária a modelagem através de distribuições com suporte contínuo positivo.

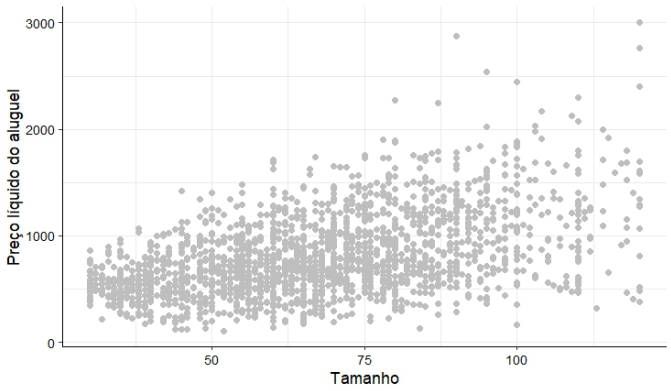

Figura 2: Gráfico de dispersão entre o preço do aluguel e o tamanho das acomodações.

Inicialmente, foi ajustado um modelo de regressão normal, utilizando a variável tamanho para modelar a média da variável aluguel (denominado de Modelo 1). A estrutura da regressão é dada por

$$Y_i \overset{\text{ind}}{\sim} \text{NO}(\mu_i, \sigma)$$

$$\mu_i = \beta_0 + \beta_1 X_i$$

em que Y é a variável resposta aluguel, X é a variável explicativa tamanho, e β_0 e β_1 são os coeficientes de regressão.

A fim de avaliar a qualidade do ajuste do Modelo 1 aos dados, foram feitos gráficos de diagnóstico baseados no resíduo quantílico a partir da função *plot()* do R, apresentados na Figura 3.

A partir do gráfico da Figura 3(a), observa-se que conforme os valores ajustados de μ aumentam, a variância dos resíduos também aumenta, indicando uma inadequabilidade do modelo proposto. Esse comportamento dos resíduos pode ser um indicativo também de que seja necessário o ajuste de um modelo que acomode heterocedasticidade na variável resposta. A partir da Figura 3(b), aparenta-se não haver dependência entre os resíduos e os índices das observações. Por fim, a partir das Figuras 3(c) e 3(d), observam-se desvios da normalidade dos resíduos devido à assimetria da densidade estimada e à presença de vários pontos divergentes da linha de tendência do *QQ-plot*, indicando desvios da suposição de normalidade da resposta e, portanto, indicando a inadequabilidade do Modelo 1. Em sequência, foi feito o *worm plot* para investigar mais a fundo a qualidade do ajuste do Modelo 1 aos dados, apresentado na Figura 4.

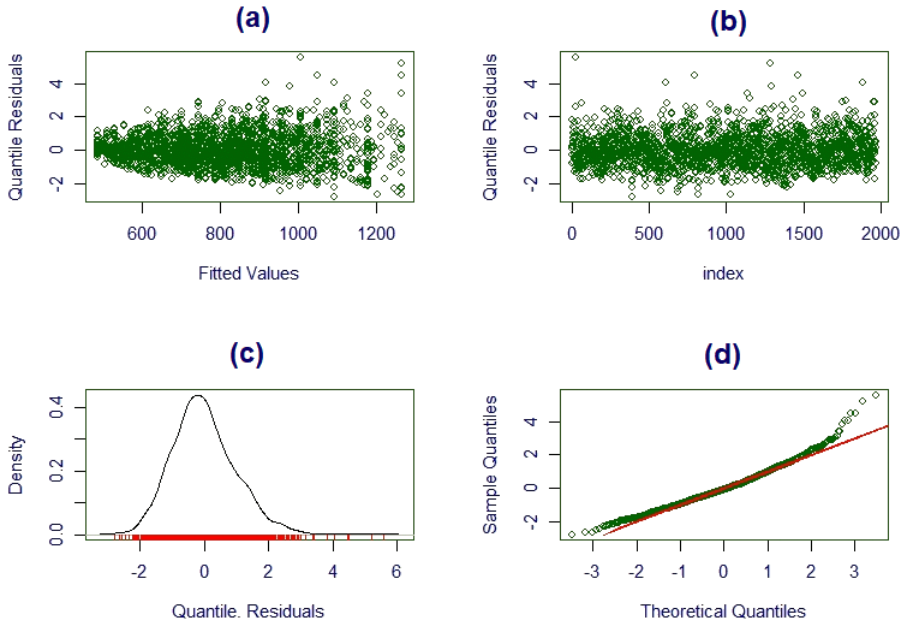

Figura 3: Gráficos de diagnóstico para o Modelo 1: (a) dispersão dos resíduos pelos valores ajustados de μ; (b) dispersão dos resíduos pelos índices das observações; (c) densidade estimada dos resíduos; (d) *QQ-plot* dos resíduos.

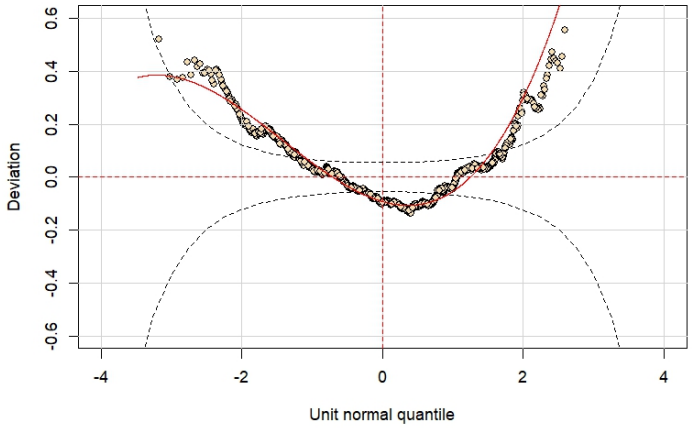

Figura 4: *Worm plot* do ajuste do Modelo 1.

A partir da Figura 4, percebe-se que o ajuste do Modelo 1 não foi adequado aos dados, uma vez que o *worm plot* apresenta forma de U e vários pontos caem fora das bandas de confiança. A forma de U do *worm plot* indica a necessidade de ajustar uma distribuição com assimetria à direita. Assim, em seguida, foram ajustados MLGs utilizando distribuição gama aos dados.

Sob a distribuição gama, foram considerados 3 modelos diferentes, com funções de ligação inversa (ligação canônica da distribuição gama, denominado Modelo 2), lo-

garítmica (denominado Modelo 3) e identidade (denominado Modelo 4). A estrutura da regressão para esses modelos é dada por

$$Y_i \overset{\text{ind}}{\sim} \text{GA}(\mu_i, \sigma)$$

$$g(\mu_i) = \beta_0 + \beta_1 X_i$$

em que Y é a variável resposta aluguel, X é a variável explicativa tamanho, β_0 e β_1 são os parâmetros de regressão, e $g(\cdot)$ é a função ligação do modelo, em que essa é a função de ligação inversa, logarítmica e identidade para os Modelos 2, 3, e 4, respectivamente. Nota-se que no modelo de RLN, enquanto a variância é constante para diferentes valores da covariável e da média, no modelo com distribuição gama essa variância aumenta conforme a variável tamanho aumenta (e consequentemente a média aumenta, como discutido na Subseção 2.7.2), o que parece se adequar mais ao comportamento dos dados (ver Figura 2).

Para avaliar a qualidade do ajuste aos dados sob os MLGs gama, foram feitos os gráficos de diagnóstico baseados no resíduo quantílico, os quais são apresentados na Figura 5 para o Modelo 2, na Figura 6 para o Modelo 3, e na Figura 7 para o Modelo 4.

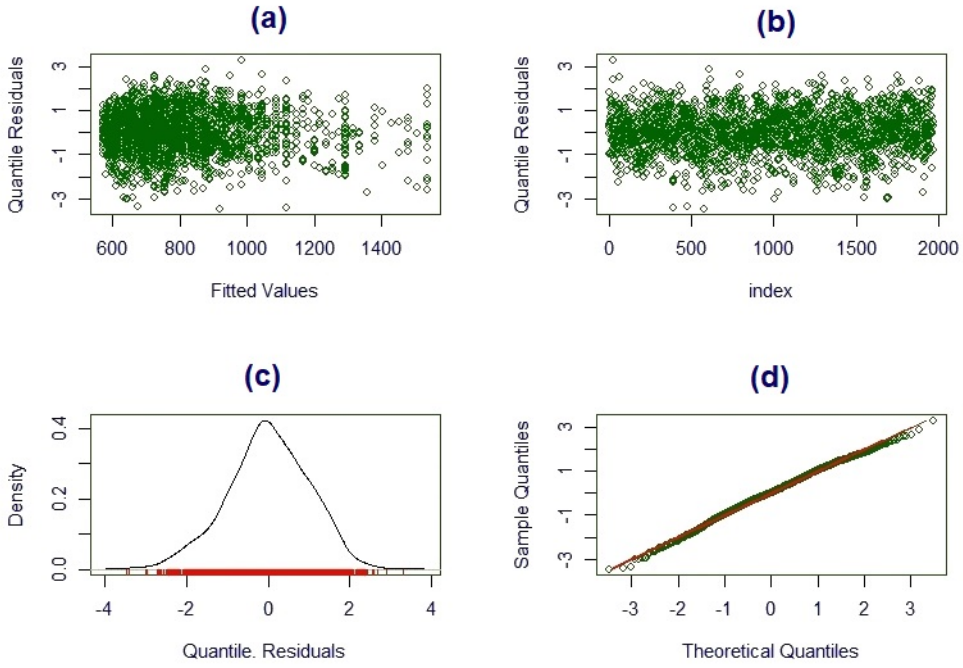

Figura 5: Gráficos de diagnóstico para o Modelo 2: (a) dispersão dos resíduos pelos valores ajustados de μ; (b) dispersão dos resíduos pelos índices das observações; (c) densidade estimada dos resíduos; (d) *QQ-plot* dos resíduos.

A partir da Figura 5(a) referente ao Modelo 2, nota-se que não há mais com-

portamento heterocedástico nos resíduos quando comparados com os valores ajustados de μ. Os resíduos ainda aparentam ter comportamento aleatório em torno de zero como se observa nas Figuras 5(a) e 5(b), sem nenhuma relação aos valores ajustados de μ ou aos índices das observações, indicando não correlação entre esses e os resíduos. Contudo, ao analisar as Figuras 5(c) e 5(d), percebem-se desvios da distribuição observada dos resíduos à esperada (a qual sob ajuste adequado do modelo é assintoticamente normal padrão), devido à assimetria à esquerda da densidade estimada e à presença de pontos divergentes da linha de tendência do *QQ-plot*. Assim, aparentam haver desvios das suposições feitas sob o Modelo 2.

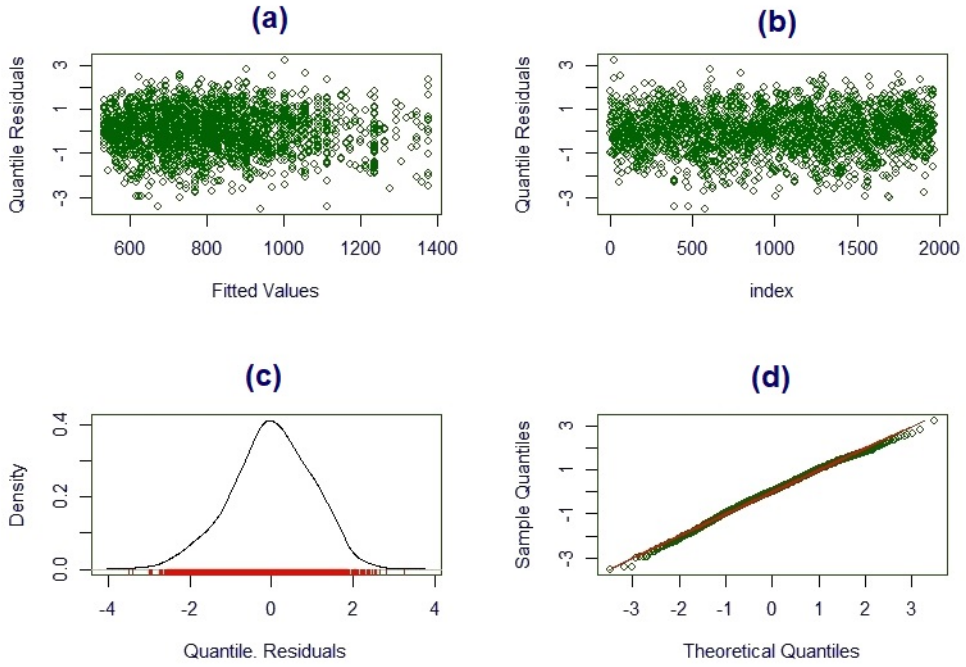

Figura 6: Gráficos de diagnóstico para o Modelo 3: (a) dispersão dos resíduos pelos valores ajustados de μ; (b) dispersão dos resíduos pelos índices das observações; (c) densidade estimada dos resíduos; (d) *QQ-plot* dos resíduos.

Os gráficos de diagnóstico para o Modelo 3 ilustrados na Figura 6 apresentam resultados similares aos obtidos para o Modelo 2, onde os resíduos parecem ser homocedásticos e não correlacionados aos valores ajustados de μ e das observações, entretanto o comportamento desses desviam da normalidade devido à presença de assimetria à esquerda.

A Figura 7 mostra que para o Modelo 4, assim como para os demais modelos ajustados a partir da distribuição gama, os resíduos aparentam ser homocedásticos e não correlacionados aos valores ajustados de μ e das observações, porém apresentam comportamento divergente da normalidade devido à assimetria à esquerda.

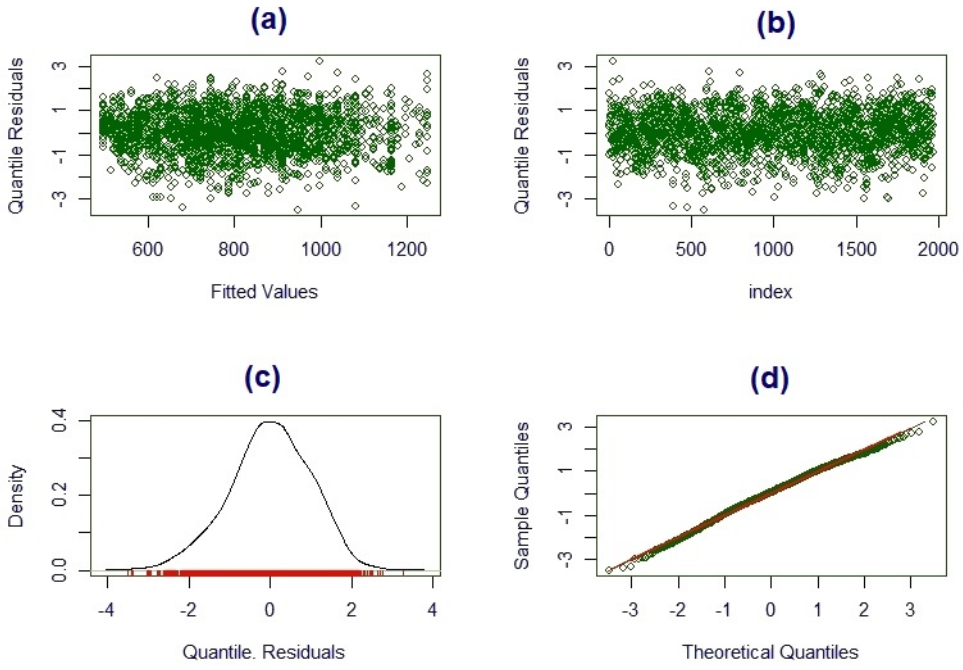

Figura 7: Gráficos de diagnóstico para o Modelo 4: (a) dispersão dos resíduos pelos valores ajustados de μ; (b) dispersão dos resíduos pelos índices das observações; (c) densidade estimada dos resíduos; (d) *QQ-plot* dos resíduos.

Em sequência, foram feitos *worm plots* para investigar com maior detalhe a qualidade do ajuste dos Modelos 2, 3 e 4 aos dados, apresentados na Figura 8.

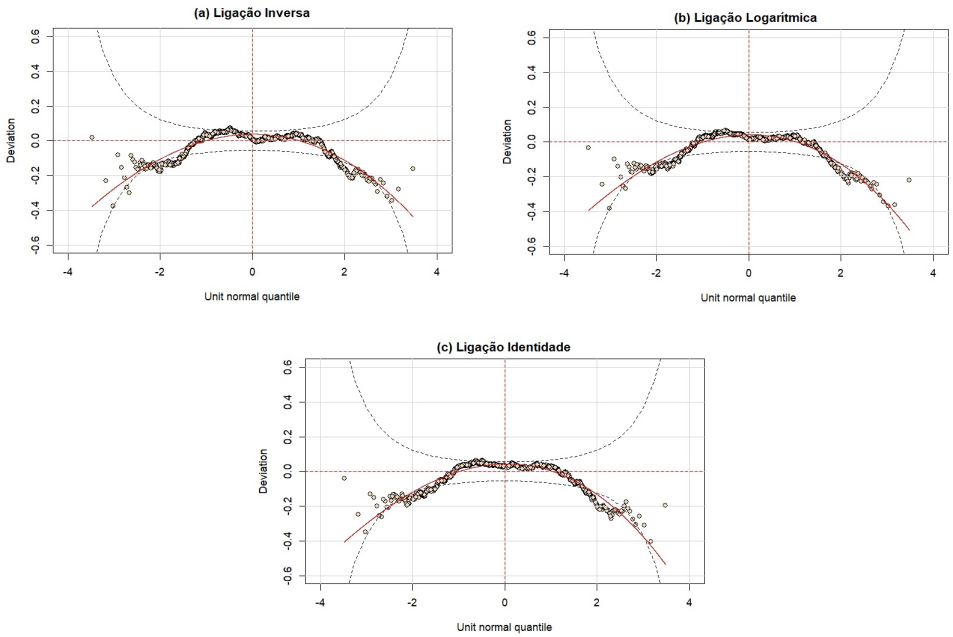

Figura 8: *Worm plots* dos ajustes dos modelos: (a) Modelo 2; (b) Modelo 3; (c) Modelo 4.

A partir dos gráficos apresentados na Figura 8, percebe-se que o ajuste do modelo de regressão gama não foi adequado aos dados para qualquer função de ligação usada, uma vez que os respectivos *worm plots* apresentam forma de U invertido e vários pontos estão fora das bandas de confiança. A forma de U invertido no gráfico indica a necessidade de ajustar uma distribuição de probabilidade à resposta com menos assimetria à direita. Necessita-se, assim, de uma distribuição com assimetria à direita, porém com menos assimetria do que a distribuição gama. Assim, foi ajustado um GAMLSS utilizando a distribuição Box-Cox Cole e Green aos dados.

A partir da distribuição Box-Cox Cole e Green, foram testados modelos com função de ligação identidade (denominado Modelo 5) e função de ligação logarítmica (denominado Modelo 6), a fim de ser possível fazer uma comparação aos modelos ajustados anteriormente. A estrutura da regressão para esses modelos é dada por

$$Y_i \overset{\text{ind}}{\sim} \text{BCCG}(\mu_i, \sigma, \nu)$$

$$g(\mu_i) = \beta_0 + \beta_1 X_i$$

em que Y é a variável resposta aluguel, X é a variável explicativa tamanho, β_0 e β_1 são os coeficientes de regressão associados ao parâmetro μ_i (aproximadamente a mediana da distribuição BCCG), e $g(\cdot)$ é a função ligação do modelo, em que se considera ligação identidade e logarítmica para os Modelos 5 e 6, respectivamente.

Assim como na modelagem pela distribuição gama, modelos ajustados pela distribuição Box-Cox Cole e Green conseguem acomodar dados heterocedásticos. Contudo, diferente dos modelos de regressão gama, esses conseguem acomodar diversos comportamentos diferentes de assimetria da variável resposta. Assim, a partir do parâmetro ν da distribuição BCCG, espera-se comportar adequadamente a assimetria dos dados.

Para avaliar a qualidade dos ajustes do Modelo 5 e do Modelo 6 aos dados, foram elaborados gráficos de diagnóstico baseados no resíduo quantílico, apresentados na Figura 9 e na Figura 10.

A partir das Figuras 9(a) e 9(b), observa-se que os resíduos quantílicos apresentam comportamento aleatório ao redor de zero perante toda a faixa de valores ajustados para μ, além de não apresentarem nenhuma tendência em relação aos índices das observações. Esse comportamento indica a homocedasticidade dos resíduos e a não correlação destes com os valores ajustados de μ e com as observações. Além disso, o gráfico da densidade estimada dos resíduos se assemelha à densidade de uma distribuição normal e o *QQ-plot* não apresenta nenhum ponto ou comportamento afastado da linha de tendência,

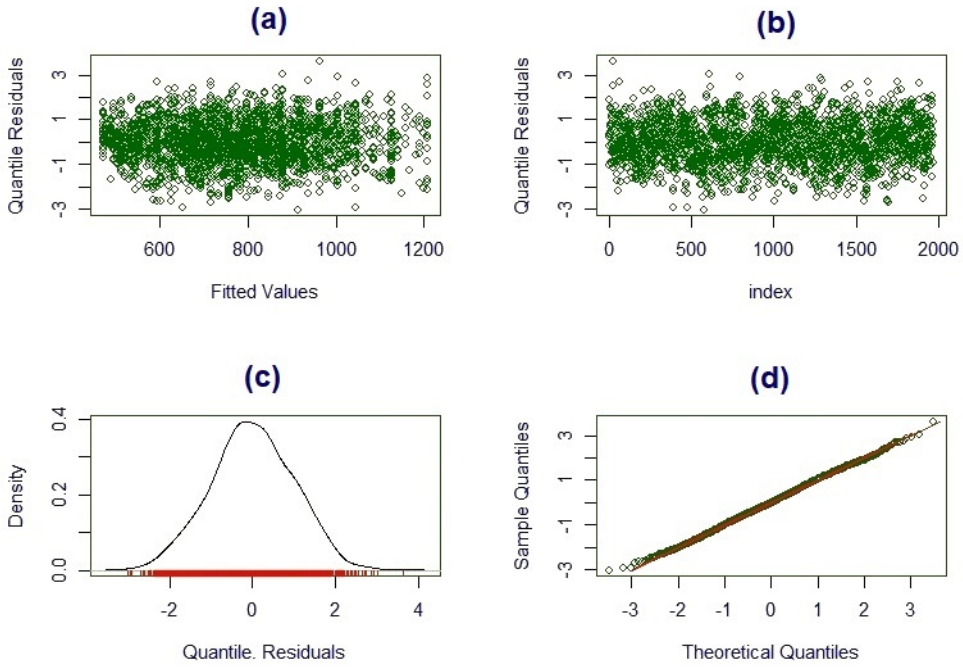

Figura 9: Gráficos de diagnóstico para o Modelo 5: (a) dispersão dos resíduos pelos valores ajustados de μ; (b) dispersão dos resíduos pelos índices das observações; (c) densidade estimada dos resíduos; (d) *QQ-plot* dos resíduos.

indicando a normalidade dos resíduos quantílicos. A partir dos resultados desses gráficos, não aparenta haver nenhum desvio dos pressupostos do modelo, e não há evidências de que o Modelo 5 não se ajustou adequadamente ao comportamento dos dados.

Os gráficos esboçados na Figura 10 indicam que os resíduos do Modelo 6 tiveram comportamento similar aos resíduos do Modelo 5. Quando utilizando a distribuição BCCG e função de ligação logarítmica, os resíduos também apresentaram comportamento adequado às suposições do modelo. Portanto, análogo aos resultados sob o Modelo 5, não há indícios apontando que o Modelo 6 não se ajustou adequadamente aos dados.

Em sequência, foram feitos os *worm plots* para investigar mais detalhadamente a qualidade do ajuste dos Modelos 5 e 6 aos dados, apresentado na Figura 11. A partir dos gráficos, percebe-se que mais de 95% dos dados estão dentro das bandas de confiança em ambos *worm plots*, e os pontos não possuem nenhum padrão fora do comum para ambos modelos. Assim, parece que os modelos de regressão BCCG supostos aos dados estão adequados.

A Tabela 7 apresenta as estimativas de máxima verossimilhança dos coeficientes dos modelos, além de critérios de informação para os modelos, a fim de ser possível comparar os diferentes modelos ajustados. Ressalta-se que os critérios de informação não são muito adequados para comparar modelos não encaixados, mas podem ser utilizados

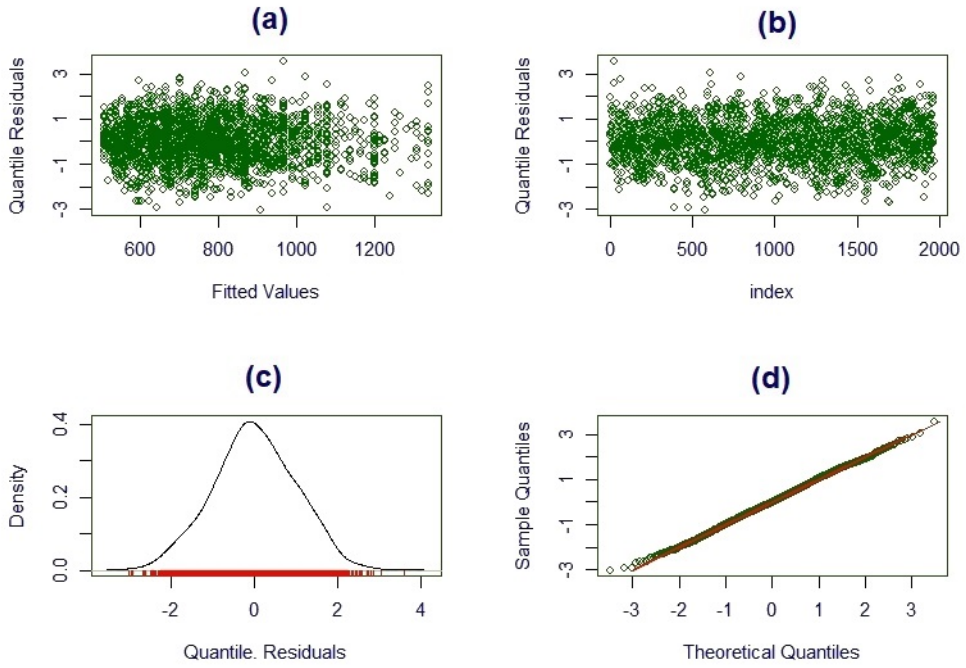

Figura 10: Gráficos de diagnóstico para o Modelo 6: (a) dispersão dos resíduos pelos valores ajustados de μ; (b) dispersão dos resíduos pelos índices das observações; (c) densidade estimada dos resíduos; (d) *QQ-plot* dos resíduos.

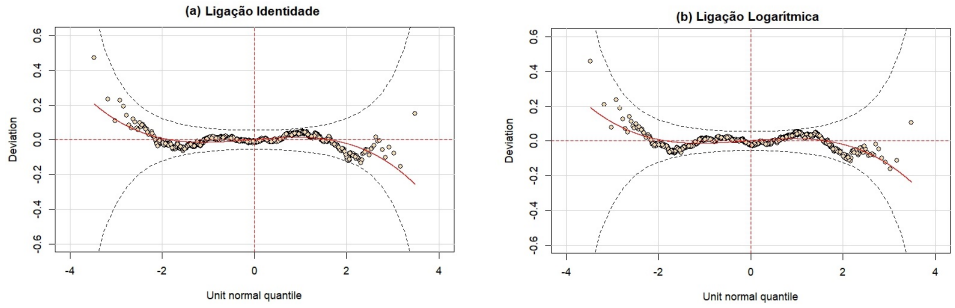

Figura 11: *Worm plot* dos ajustes dos modelos: (a) Modelo 5; (b) Modelo 6.

com as devidas ressalvas.

Pela Tabela 7, percebe-se que o modelo RLN apresentou os maiores valores dos critérios de informação entre os modelos ajustados, resultado este que está de acordo com a análise diagnóstica feita, indicando para a inferioridade desse modelo quando comparado com os demais. Entre os modelos de regressão gama, o modelo ajustado com função de ligação inversa apresentou os maiores valores dos critérios de informação. Esse resultado aponta que mesmo essa sendo a função de ligação canônica para distribuição gama, são preferíveis os ajustes a partir das demais funções de ligação. Os diagnósticos para ambos modelos ajustados pela distribuição BCCG apontaram que esses apresentam ajustes adequados aos dados. Dessa forma, diferentes critérios devem ser implementados para a

Tabela 7: Modelos ajustados na Aplicação 1, suas estimativas e valores para critérios de informação.

Modelo	Distribuição	Função de Ligação para μ	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\sigma}$	$\hat{\nu}$	AIC	BIC	GAIC(2, 5)
Modelo 1	NO	Identidade	226,19	8,65	333,26	—	28469,35	28486,11	28470,85
Modelo 2	GA	Inversa	0,0021	-0,000012	0,4078	—	28123,58	28140,34	28125,08
Modelo 3	GA	Logarítmica	5,96	0,011	0,4060	—	28104,43	28121,18	28105,93
Modelo 4	GA	Identidade	248,06	8,31	0,4062	—	28106,93	28123,68	28108,43
Modelo 5	BCCG	Identidade	229,08	8,15	0,4057	0,52	28089,88	28112,22	28091,88
Modelo 6	BCCG	Logarítmica	5,91	0,011	0,4059	0,51	28088,85	28111,19	28090,85

escolha do melhor modelo entre esses. Informações prévias ou um conhecimento teórico da relação esperada dos preços de aluguel com o tamanho de acomodações podem apontar para qual modelo é mais adequado aos dados. Entretanto, utilizando apenas princípios estatísticos, nota-se que os critérios de informação para o modelo com ligação logarítmica apresentaram valores melhores que os critérios para o modelo com ligação identidade, sendo então preferível o Modelo 6.

Para as distribuições NO, GA e BCCG, o parâmetro μ apresenta interpretação similar, sendo uma medida de tendência central (aproximadamente para a distribuição BCCG). Assim, quando utilizando a mesma função de ligação, modelos de regressão ajustados sob essas distribuições podem apresentar valores próximos. Ao analisar os valores de $\hat{\beta}_0$ e $\hat{\beta}_1$ para os modelos ajustados com função de ligação identidade (Modelo 1, Modelo 4 e Modelo 5), percebe-se que as estimativas apresentam valores próximos entre si. Entretanto, pelo Modelo 5 não ter apresentado nenhuma inadequabilidade no diagnóstico, espera-se que os estimadores por esse modelo apresentem menor erro padrão e, portanto, sejam mais precisos. Assim, dado o ajuste adequado do Modelo 5, a interpretação dos parâmetros é confiável. Por esse modelo, interpreta-se $\hat{\beta}_1 = 8,15$ como aproximadamente o aumento na mediana do preço do aluguel causado pelo aumento de um metro quadrado na área da acomodação.

Para os modelos ajustados utilizando função de ligação logarítmica (Modelo 3 e Modelo 6), também se observam valores ajustados dos coeficientes de regressão próximos entre os dois modelos. Uma vez que o ajuste pelo Modelo 6 não apresentou nenhuma inadequabilidade no diagnóstico, espera-se que os estimadores por esse modelo apresentem menor erro padrão e, assim, sejam mais precisos. Sendo o ajuste pelo Modelo 6 adequado aos dados, a interpretação dos parâmetros é confiável. Por esse modelo, $\hat{\beta}_1 = 0,011$ e, assim, interpreta-se que o aumento de um metro quadrado na área da acomodação causa um aumento aproximado de 1,1% na mediana do preço do aluguel.

Nota-se que para todos modelos, exceto o Modelo 1, os valores de $\hat{\sigma}$ são similares. Isso ocorre uma vez que a interpretação de σ é diferente na distribuição NO e na distribuição GA e BCCG, sendo na distribuição NO a interpretação dada pelo desvio padrão

da distribuição da variável resposta, e nas distribuições GA e BCCG, esse valor representa o coeficiente de variação da distribuição (aproximadamente para a distribuição BCCG). Assim, os modelos ajustados sob distribuição Gama e sob distribuição Box-Cox Cole e Green acomodam o comportamento heterocedástico da variável resposta, uma vez que por esses modelos o acréscimo da variável explicativa causa um acréscimo no desvio padrão da variável resposta, enquanto que o modelo ajustado sob distribuição normal assume homocedasticidade da variável resposta. Esse cenário ilustra que mesmo a modelagem por RLN sendo vastamente utilizada, esses modelos podem gerar ajustes errôneos aos dados, e levar a conclusões equivocadas sobre o comportamento da variável resposta. Sendo os ajustes pelo Modelo 5 ou pelo Modelo 6 (baseados na distribuição BCCG) adequados aos dados, espera-se que os estimadores $\hat{\sigma}$ desses modelos sejam mais precisos que nos demais modelos.

4.2 Aplicação 2: Número de canais de TV públicos de boa qualidade disponíveis por características de áreas metropolitanas

A modelagem por regressão do número de canais públicos disponíveis por regiões metropolitanas pode auxiliar na identificação de fatores que aumentam ou diminuem esse número, e pode servir de base para a criação de políticas públicas voltadas a esse assunto ou critérios de investimentos de empresas de mídias ou publicidades conforme as demandas regionais. A partir do banco de dados referente a demandas de TV em 40 regiões metropolitanas dos Estados Unidos (Ramanathan, 1993), espera-se ajustar um modelo de regressão paramétrico adequado para explicar o número de canais públicos disponíveis a partir do número de assinantes de TV a cabo, número de domicílios, renda per capita domiciliar, e número de canais a cabo disponíveis nas áreas.

A Figura 12 apresenta os gráficos de dispersão do número de canais públicos disponíveis pelas demais covariáveis. A partir dos gráficos de dispersão, percebe-se que os dados da variável resposta são dados discretos estritamente positivos, indicando que essa variável seja referente a dados de contagem. Também, observa-se que a variável resposta apresenta maior concentração entre valores de quatro a oito, com alguns valores acima de oito, mas nenhum valor abaixo de quatro, indicando que o ajuste da variável resposta por uma distribuição com moda tendendo a zero pode não ser adequado. A partir da Figura 12(a), observa-se no geral uma tendência de crescimento para o número de canais públicos conforme o aumento do número de assinantes de TV a cabo. Nesse gráfico, nota-se também a presença de uma observação com valor alto para o número de

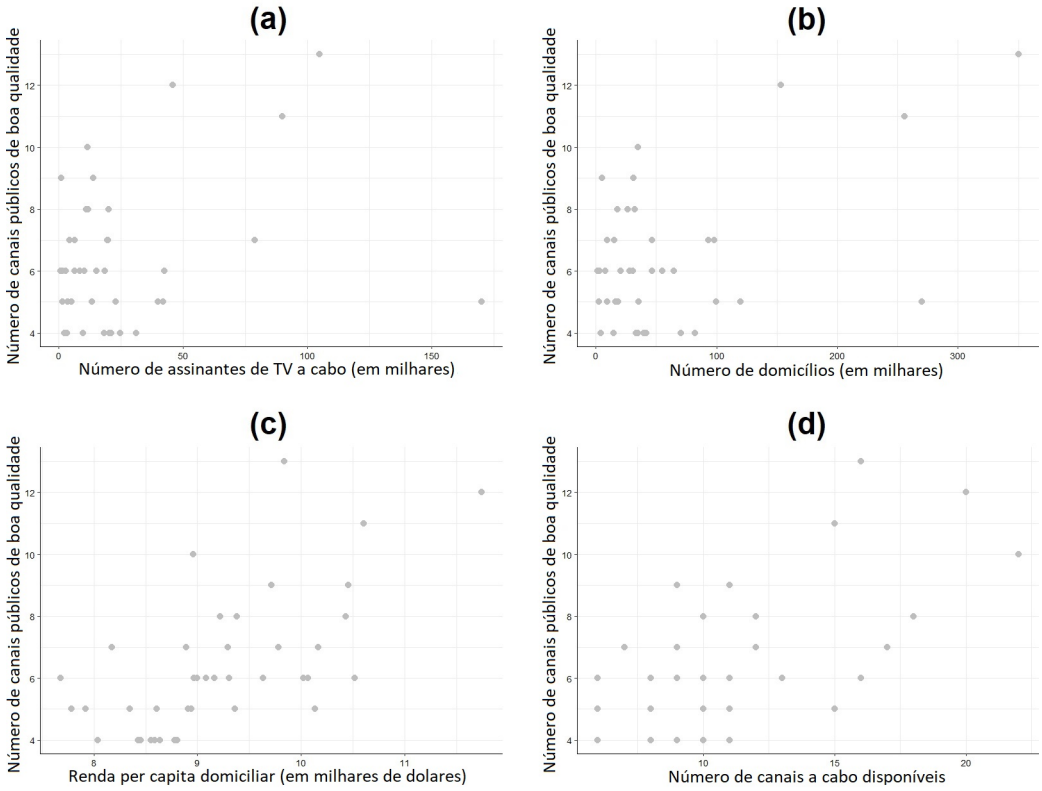

Figura 12: Gráficos de dispersão entre o número de canais públicos disponíveis e (a) o número de assinantes de TV a cabo; (b) o número de domicílios; (c) a renda per capita domiciliar; (d) o número de canais a cabo disponíveis.

assinantes, entretanto baixo para o número de canais públicos, que pode vir a ser um ponto influente ou *outlier* no ajuste dos modelos. Os demais gráficos também apresentam comportamento similar, onde há uma tendência de crescimento para o número de canais públicos com o acréscimo do número de domicílios, o acréscimo da renda per capita e o acréscimo do número de canais a cabo disponíveis. Na Figura 12(c), nota-se também uma observação com valor alto do número de domicílios, mas valor baixo para o número de canais públicos, a qual pode ser uma observação influente no ajuste ou *outlier*.

A Tabela 8 apresenta as distribuições discretas de contagem ajustadas marginalmente à variável resposta, além do valor do GAIC($\kappa = 2, 5$) para esses ajustes.

Tabela 8: Valores do critério de informação GAIC(2,5) para diferentes distribuições discretas de contagem ajustadas marginalmente para o número de canais públicos disponíveis.

Distribuição	GAIC(2,5)	Distribuição	GAIC(2,5)	Distribuição	GAIC(2,5)	Distribuição	GAIC(2,5)
DPO	177,0771	ZIP2	179,6966	DEL	182,1966	LG	260,4264
PO	177,1966	PIG	179,6966	BNB	182,1966	ZALG	262,9264
ZAP	179,5594	ZANBI	182,0594	ZINBF	184,6966	ZIPF	294,4928
GPO	179,6966	ZAPIG	182,0595	ZIBNB	184,6966	ZAZIPF	296,9928
NBI	179,6966	NBF	182,1966	GEOM	236,6607	YULE	326,7639
NBII	179,6966	ZINBI	182,1966	GEOMo	236,6607		
ZIP	179,6966	ZIPIG	182,1966	WARING	239,1607		

A partir da Tabela 8, pode-se ter uma intuição inicial de quais distribuições de probabilidade podem ser apropriadas para a modelagem da variável resposta. Observa-se que marginalmente a distribuição Poisson apresentou o segundo melhor valor para o critério de informação entre as distribuições testadas. Por essa distribuição ser da família exponencial uniparamétrica, foi inicialmente ajustado um MLG sob a distribuição Poisson (denominado de Modelo 11), com estrutura da regressão dada por

$$Y_i \overset{\text{ind}}{\sim} \text{PO}(\mu_i)$$

$$g(\mu_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i}$$

em que Y é a variável resposta número de canais públicos, X_1, X_2, X_3, e X_4 são as variáveis explicativas número de assinantes de TV a cabo, número de domicílios, renda per capita domiciliar, e número de canais a cabo, respectivamente, β_0, β_1, β_2 β_3 e β_4 são os parâmetros regressores, e $g(\cdot)$ é a função ligação do modelo, sendo essa a função de ligação logarítmica para o modelo proposto.

Para avaliar a qualidade do ajuste do modelo a partir da distribuição Poisson, foram calculados os resíduos quantílicos e construídos os gráficos de diagnóstico para o modelo, esboçados na Figura 13.

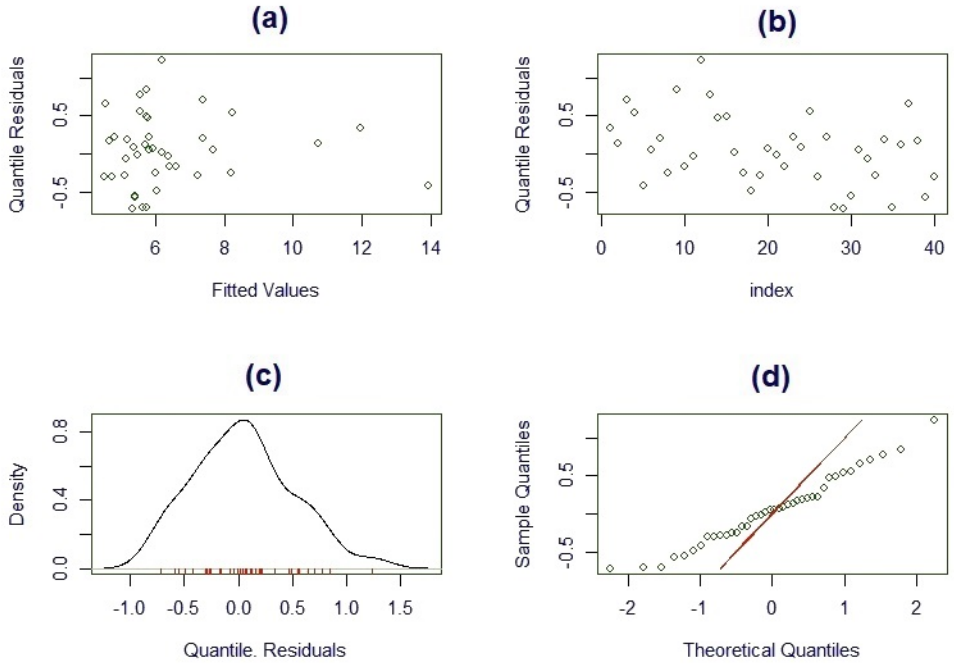

Figura 13: Gráficos de diagnóstico para o Modelo 11: (a) dispersão dos resíduos pelos valores ajustados de μ; (b) dispersão dos resíduos pelos índices das observações; (c) densidade estimada dos resíduos; (d) *QQ-plot* dos resíduos.

A partir da Figura 13(a), observa-se que a maioria dos valores ajustados para μ está entre quatro e oito, com alguns valores mais elevados, como esperado de um bom ajuste dado o comportamento observado da variável resposta. Ainda, analisando a Figura 13(a), verifica-se que a variância dos resíduos para os valores ajustados de μ mais elevados parece ser menor de que para valores ajustado menores. Visto que o resíduo quantílico sob modelos discretos é aleatorizado, os resíduos foram recalculados várias vezes, onde esse comportamento persistiu, indicando uma inadequabilidade no ajuste. Entretanto vale notar que esse comportamento também pode ter ocorrido devido ao tamanho pequeno da amostra. A Figura 13(b) não indica nenhuma tendência clara, indicando não correlação dos resíduos aos índices das observações. Analisando as Figuras 13(c) e 13(d), observa-se que os resíduos apresentam grandes divergências ao comportamento esperado de uma distribuição normal, principalmente devido aos pontos do *QQ-plot* não seguirem a linha de tendência, assim indicando inadequabilidade do ajuste. Vale notar que esse comportamento também persistiu quando os resíduos quantílicos foram recalculados.

Em sequência, foi feito o *worm plot* para o Modelo 11, apresentado na Figura 14, a fim de se compreender mais detalhadamente o ajuste do modelo. A partir do *worm plot*, nota-se uma clara tendência no comportamento dos dados, e diversos pontos fora das bandas de confiança do gráfico, indicando que o ajuste do Modelo 11 não foi adequado aos dados. Pelo comportamento linear no gráfico, com inclinação negativa, é indicada a necessidade de ajustar um modelo a partir de uma distribuição com menos variância que a distribuição Poisson, ou seja, com subdispersão em relação à distribuição Poisson. Assim, foi feito um ajuste de um GAMLSS a partir da distribuição Poisson dupla.

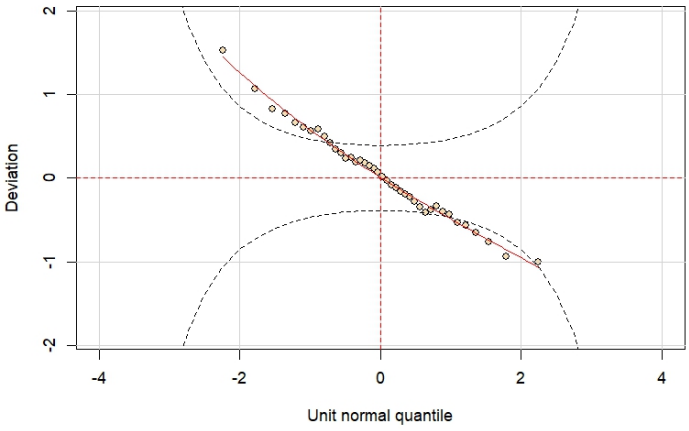

Figura 14: *Worm plot* do ajuste do Modelo 11.

Foi ajustado um GAMLSS sob a distribuição Poisson dupla (denominado de Modelo 12), com estrutura da regressão dada por

$$Y_i \overset{\text{ind}}{\sim} \text{DPO}(\mu_i, \sigma)$$

$$g(\mu_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i}$$

em que Y é a variável resposta número de canais públicos, X_1, X_2, X_3, e X_4 são os valores fixados das covariáveis número de assinantes de TV a cabo, número de domicílios, renda per capita domiciliar, e número de canais a cabo, respectivamente, β_0, β_1, β_2 β_3 e β_4 são os coeficientes de regressão associados ao parâmetro μ_i (aproximadamente a média da distribuição DPO), e $g(\cdot)$ é a função ligação do modelo, sendo essa a função de ligação logarítmica para o modelo proposto. Nota-se que não foi incorporada estrutura de regressão para o parâmetro σ e, portanto, esse se manterá constante para todas observações.

A fim de se avaliar a qualidade do ajuste do modelo a partir da distribuição Poisson dupla, foram calculados os resíduos quantílicos e apresentados gráficos de diagnóstico para o modelo, esboçados na Figura 15.

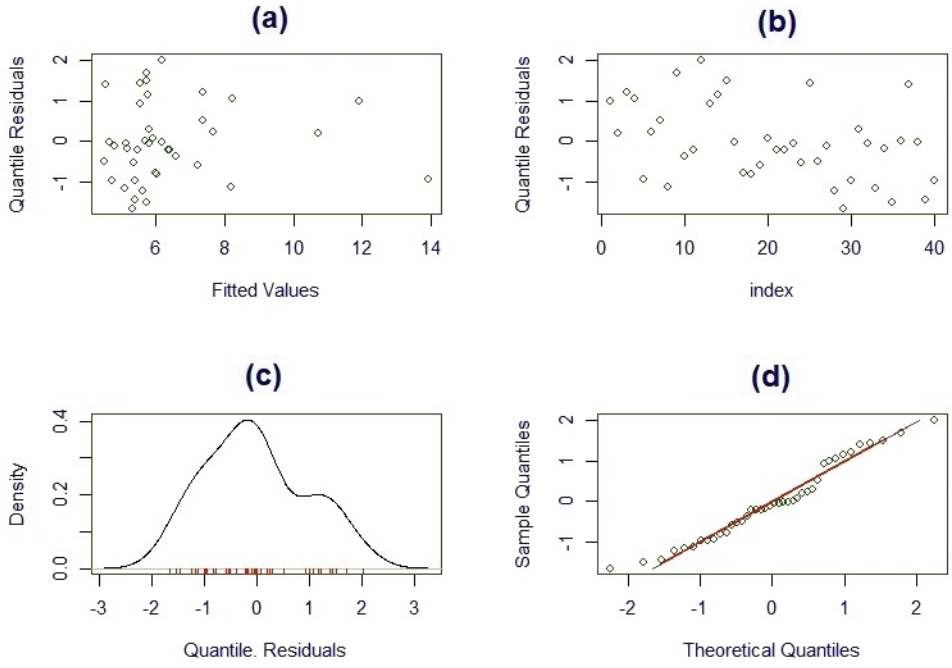

Figura 15: Gráficos de diagnóstico para o Modelo 12: (a) dispersão dos resíduos pelos valores ajustados de μ; (b) dispersão dos resíduos pelos índices das observações; (c) densidade estimada dos resíduos; (d) *QQ-plot* dos resíduos.

A partir da Figura 15(a), observa-se que a maioria dos valores ajustados para μ está entre quatro e oito, com algumas observações com valores mais elevados, um

comportamento condizente com os valores observados da variável resposta. Também, o comportamento heterocedástico dos resíduos observado no ajuste do Modelo 11 parece ter sido amenizado no ajuste do Modelo 12. Esse comportamento, que se repetiu para diferentes cálculos dos resíduos quantílicos, indica adequação do modelo aos dados. A partir da Figura 15(b), não se observa nenhuma tendência evidente, indicando não correlação entre os índices das observações e os resíduos. Embora o gráfico da densidade estimada dos resíduos não se assemelhe à densidade da distribuição normal, isso se deve ao número baixo de observações no banco, e o comportamento apresentado no *QQ-Plot* indica a normalidade dos resíduos. Assim, não há nenhuma quebra das suposições do modelo e, portanto, não há indícios de que o ajuste do modelo esteja inadequado.

Por fim, foi feito o *worm plot* para o Modelo 12 a fim de verificar mais detalhadamente o ajuste do modelo, apresentado na Figura 16. Uma vez que os pontos estão próximos de zero, não há nenhum padrão distinguível no gráfico, e os pontos estão dentro das bandas de confiança, não há indícios contra a adequação do modelo de regressão Poisson dupla aos dados.

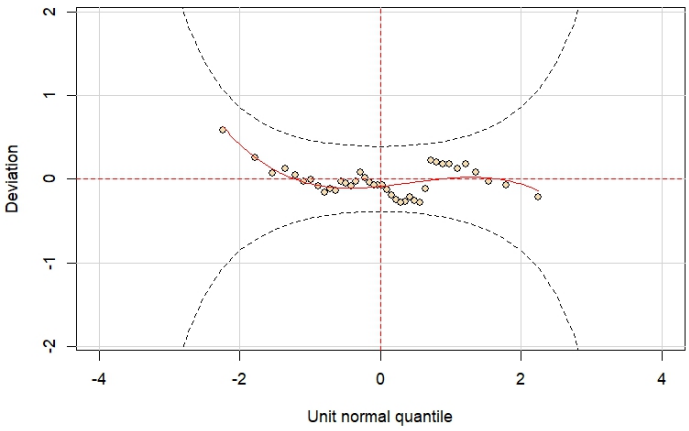

Figura 16: *Worm plot* do ajuste do Modelo 12.

Uma vez identificada a distribuição adequada aos dados, é necessária a seleção das variáveis do modelo final. Como foi introduzida uma estrutura de regressão apenas para o parâmetro μ da distribuição Poisson dupla, a seleção das variáveis será feita apenas para esse parâmetro. A partir do Modelo 12, foram ajustados os Modelo 13, 14, 15, e 16, sendo retiradas da estrutura de regressão as variáveis número de assinantes de TV a cabo, número de domicílios, renda per capita domiciliar, e número de canais a cabo, respectivamente em cada modelo. Assim, será feita a seleção de variáveis a partir do critério de seleção *backward*, usando o GAIC(2,5) como critério de decisão entre modelos.

A Tabela 9 apresenta os modelos ajustados e suas medidas para diferentes critérios de informação. Vale notar que o Modelo 12 generaliza todos os demais modelos apresentados na aplicação, assim sendo possível comparar os valores dos critérios de informação de todos os modelos.

Tabela 9: Modelos ajustados na Aplicação 2, e suas medidas para critérios de informação.

Modelo	Distribuição	AIC	BIC	GAIC$(2,5)$
Modelo 11	PO	165,05	173,50	167,55
Modelo 12	DPO	135,26	145,40	138,26
Modelo 13	DPO	140,96	149,41	143,46
Modelo 14	DPO	144,95	153,39	147,45
Modelo 15	DPO	149,48	157,93	151,98
Modelo 16	DPO	144,40	152,85	146,90

A partir da Tabela 9, é perceptível que o Modelo 12 apresentou melhores resultados ao Modelo 11 em relação a todos os critérios de informação analisados. Assim, a distribuição DPO é preferível à distribuição PO pelos critérios de informação. Também, os modelos retirando uma variável em relação ao Modelo 12 apresentaram resultados piores de que o Modelo 12 nos critérios de informação, sendo assim preferível o modelo que mantém todas as variáveis na estrutura de regressão.

A fim de comparar os valores ajustados para cada modelo, as estimativas dos coeficientes para cada modelo são apresentadas na Tabela 10. Denota-se $\hat{\beta}_{\mu 0}$, $\hat{\beta}_{\mu 1}$, $\hat{\beta}_{\mu 2}$, $\hat{\beta}_{\mu 3}$, $\hat{\beta}_{\mu 4}$ os coeficientes de regressão para o parâmetro μ ajustados referentes ao intercepto, ao número de assinantes de TV a cabo, ao número de domicílios, à renda per capita domiciliar, e ao número de canais a cabo respectivamente.

Tabela 10: Modelos ajustados na Aplicação 2, e suas estimativas.

Modelo	Distribuição	$\hat{\beta}_{\mu 0}$	$\hat{\beta}_{\mu 1}$	$\hat{\beta}_{\mu 2}$	$\hat{\beta}_{\mu 3}$	$\hat{\beta}_{\mu 4}$	$\log(\hat{\sigma})$
Modelo 11	PO	0,0816	-0,0057	0,0030	0,1520	0,0286	—
Modelo 12	DPO	0,0796	-0,0057	0,0030	0,1521	0,0286	-1,57
Modelo 13	DPO	-0,0681	—	0,0009	0,1666	0,0285	-1,38
Modelo 14	DPO	-0,1409	0,0007	—	0,1729	0,0333	-1,27
Modelo 15	DPO	1,3501	-0,0067	0,0035	—	0,0411	-1,17
Modelo 16	DPO	-0,0489	-0,0056	0,0035	0,1962	—	-1,29

A partir da Tabela 10, observa-se que os valores ajustados para os coeficientes no Modelo 11 e no Modelo 12 apresentam estimativas próximas. No entanto, pelo Modelo 12 ter apresentado ajuste adequado aos dados, suas estimativas são mais precisas. Quando

comparando os coeficientes no Modelo 12 com os Modelos 13, 14, 15 e 16, é notório como a remoção de certas variáveis pode causar uma mudança significativa nas estimativas dos demais coeficientes do modelo, e levar a interpretações diferentes entre os modelos. Em específico, nota-se uma grande diferença nos valores estimados de $\hat{\beta}_{\mu 1}$ no Modelo 14 e $\hat{\beta}_{\mu 2}$ no Modelo 13 quando comparados aos demais modelos. Também, percebe-se um acréscimo em $\log(\hat{\sigma})$ quando variáveis são retiradas do modelo. Uma vez que σ modela a sobredispersão (ou subdispersão) da variável resposta em relação à distribuição Poisson, esse acréscimo no parâmetro reflete um aumento na variância da variável resposta quando uma variável explicativa é retirada do modelo.

Por fim, a Tabela 11 apresenta as estimativas dos coeficientes ajustados para o Modelo 12, e as significâncias dos coeficientes a partir de testes Wald, utilizando a distribuição assintótica dos estimadores de máxima verossimilhança.

Tabela 11: Significância dos coeficientes ajustados no Modelo 12 a partir de testes Wald.

Coeficiente	Estimativa	Erro-Padrão	Estatística de Teste	p-valor
$\beta_{\mu 0}$	0,0796	0,2958	0,269	0,788
$\beta_{\mu 1}$	-0,0057	0,0020	2,828	0,005
$\beta_{\mu 2}$	0,0030	0,0008	3,740	<0,001
$\beta_{\mu 3}$	0,1521	0,0339	4,488	<0,001
$\beta_{\mu 4}$	0,0286	0,0078	3,658	<0,001
$\log(\sigma)$	-1,5656	0,2239	-6,991	<0,001

Os testes Wald referentes aos parâmetros de regressão podem auxiliar na escolha de variáveis para o modelo, uma vez que testam a significância do coeficiente de cada variável no modelo. A partir dos valores para os testes apresentados na Tabela 11, a um nível de significância de $\alpha = 0,05$, todos os coeficientes de regressão são significativamente diferentes de zero. Assim, a partir dos testes, todas as variáveis do Modelo 12 devem ser mantidas. Uma vez que a distribuição Poisson equivale à distribuição Poisson dupla com parâmetro $\sigma = 1$, o teste Wald para a significância de $\log(\sigma)$ pode ser utilizado como critério para seleção entre a distribuição PO e DPO. Como ao nível de significância de $\alpha = 0,05$ existem evidências de que $\log(\sigma)$ é significativamente diferente de zero, conclui-se que existem evidências de que σ é significativamente diferente de um, sendo então preferível o modelo ajustado sob a distribuição DPO.

Dado como adequado o ajuste do Modelo 12 aos dados, as suas estimativas são confiáveis e podem ser interpretadas. Assim, interpreta-se $\hat{\beta}_{\mu 1} = -0,0057$ como um decréscimo de 0,57% aproximadamente na média do número de canais de TV públicos

de boa qualidade disponíveis causado pelo acréscimo de mil assinantes de TV a cabo, dadas constantes as demais covariáveis. A estimativa $\hat{\beta}_{\mu 2} = 0,0030$ é interpretada como um acréscimo de 0,30% aproximadamente na média do número de canais de TV públicos disponíveis devido ao acréscimo de mil domicílios na área, dadas as demais covariáveis como constantes. O acréscimo de mil dólares na renda per capita domiciliar nas regiões metropolitanas causa um acréscimo de aproximadamente 15,21% na média do número de canais de TV públicos disponíveis, quando as demais covariáveis se mantêm constantes. Por fim, o acréscimo de um canal de TV a cabo disponível na área resulta no acréscimo de aproximadamente 2,86% na média do número de canais de TV públicos disponíveis na área, quando as demais covariáveis são mantidas constantes.

A estimativa para σ pode ser interpretada como aproximadamente o quanto de sobredispersão (ou subdispersão) a distribuição tem em relação à distribuição Poisson com a mesma média. Assim, a distribuição DPO ajustada para a variável resposta possui variância aproximadamente 79,10% menor[3] de que uma distribuição PO com a mesma média.

[3]Valor obtido a partir do cálculo $1 - e^{\log(\hat{\sigma})} = 1 - e^{-1,5656} = 1 - 0,210 = 0,7910 = 79,10\%$.

5 Conclusão

Nesse trabalho, foram discutidos e comparados os modelos de regressão linear normal, modelos lineares generalizados e os modelos aditivos generalizados para localização, escala e forma. Foi realizada uma revisão das principais distribuições contínuas com suporte real, contínuas com suporte positivo, contínuas e mistas com suporte no intervalo unitário, discretas de contagem e discretas limitadas. Em seguida, foi feita uma descrição de como várias das distribuições apresentadas se adequam à modelagem estatística via regressão, em quais situações devem ser empregadas, quais suas limitações e como são suas implementações sob os GAMLSS. A classe dos GAMLSS foi discutida em termos de suas principais aplicações, métodos inferenciais, ferramentas para seleção de modelos, técnicas de diagnóstico e limitações.

Os GAMLSS apresentam uma grande vantagem em sua aplicação para modelagem por acomodarem uma vasta gama de distribuições, além de permitirem estruturas de regressão para cada parâmetro da distribuição da variável resposta. Isso permite flexibilidade para essa classe de modelos para diversas naturezas de dados. A partir da aplicação dos GAMLSS a dados reais, foi possível entender as vantagens desses modelos sobre modelos mais simples como os MLG ou modelos de RLN. Nos dados referentes ao preço de aluguel de acomodações, foi vista uma aplicação onde um GAMLSS sob a distribuição Box-Cox Cole e Green conseguiu adequadamente modelar a mediana, o coeficiente de variação e a assimetria da variável resposta, resultando em um ajuste adequado dos dados e interpretações confiáveis dos coeficientes estimados, enquanto ajustes por MLGs e modelos RLN se mostraram inadequados. Além disso, foram apresentados alguns critérios que auxiliam na seleção de uma função de ligação adequada na modelagem. Ademais, foi ilustrado como variáveis de natureza estritamente positiva podem ser modeladas por distribuições contínuas reais, sob condição da variável resposta apresentar valores afastados de zero, e então não serem ajustados valores negativos à variável resposta. Conjuntamente, foi exemplificado como pode ser utilizada função de ligação identidade para modelar parâmetros estritamente positivos, quando os valores ajustados forem estritamente positivos e distantes de zero.

Nos dados referentes ao número de canais de TV públicos, foi vista uma aplicação onde um GAMLSS sob a distribuição Poisson dupla conseguiu adequadamente modelar a média da variável resposta e sua subdispersão em relação à distribuição Poisson. Além disso, foi exemplificado como pode ser o processo de seleção de variáveis no contexto do GAMLSS, e como critérios de informação e testes de significância podem auxiliar nessa

seleção. Foi explicada também a necessidade da avaliação dos resíduos quantílicos diversas vezes ao calculá-los para distribuições não contínuas. Em ambas aplicações, foram apresentados casos onde a troca da distribuição da variável resposta causou leves alterações nos valores estimados dos coeficientes de regressão, quando a interpretação dos coeficientes se manteve similar entre as distribuições. Destaca-se, no entanto, que as estimativas sob um modelo adequado são mais precisas do que sob modelos inadequados, e portanto devem ser buscados modelos com ajustes adequados aos dados. Na última aplicação, foi visto ademais que a inclusão de diferentes variáveis no modelo de regressão pode causar mudanças drásticas nas estimativas dos coeficientes, destacando-se a necessidade da etapa de seleção de variáveis na modelagem por regressão.

Verifica-se que além das vantagens trazidas pela flexibilidade de tais modelos, há uma grande vantagem em sua utilização devido à fácil interpretabilidade dos parâmetros e coeficientes de regressão ajustados. Ainda, verifica-se que por meio do atual trabalho, foi possível ilustrar a simplicidade da interpretação desses modelos quando utilizando estruturas de regressão estritamente paramétricas. Esse estudo pode servir como um guia preliminar de como superar as limitações dos MLGs e RLNs por meio dos GAMLSS. Assim sendo, esse trabalho pode vir a ser uma referência inicial para futuros estudos que necessitem de modelos com ajustes flexíveis aos dados, mas ainda interpretáveis.

Referências

Aeberhard, W. H. et al. Robust fitting for generalized additive models for location, scale and shape. *Statistics and Computing*, Springer, v. 31, p. 1–16, 2021.

Akaike, H. A new look at the statistical model identification. *IEEE transactions on automatic control*, Ieee, v. 19, n. 6, p. 716–723, 1974.

Akaike, H. Information measures and model selection. In: INTERNATIONAL STATISTICAL INSTITUTE. [S.l.], 1983.

Azzalini, A. A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, [Board of the Foundation of the Scandinavian Journal of Statistics, Wiley], v. 12, n. 2, p. 171–178, 1985. ISSN 03036898, 14679469. Disponível em: ⟨http://www.jstor.org/stable/4615982⟩.

Azzalini, A. Further results on a class of distributions which includes the normal ones. *Statistica*, p. 199–208, 1986.

Breslow, N. E.; Clayton, D. G. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, [American Statistical Association, Taylor Francis, Ltd.], v. 88, n. 421, p. 9–25, 1993. ISSN 01621459. Disponível em: ⟨http://www.jstor.org/stable/2290687⟩.

Buuren, S. v.; Fredriks, M. Worm plot: a simple diagnostic device for modelling growth reference curves. *Statistics in medicine*, Wiley Online Library, v. 20, n. 8, p. 1259–1277, 2001.

Buuren, S. van. Worm plot to diagnose fit in quantile regression. *Statistical Modelling*, v. 7, n. 4, p. 363–376, 2007.

Draper, N. R.; Smith, H. *Applied regression analysis*. [S.l.]: John Wiley & Sons, 1998. v. 326.

Dunn, P. K.; Smyth, G. K. Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, [American Statistical Association, Taylor Francis, Ltd., Institute of Mathematical Statistics, Interface Foundation of America], v. 5, n. 3, p. 236–244, 1996. ISSN 10618600. Disponível em: ⟨http://www.jstor.org/stable/1390802⟩.

Fahrmeir, L.; Gieger, C.; Klinger, A. Additive, dynamic and multiplicative regression. 1995.

Fernandez, C.; Steel, M. F. J. On bayesian modeling of fat tails and skewness. *Journal of the American Statistical Association*, [American Statistical Association, Taylor Francis, Ltd.], v. 93, n. 441, p. 359–371, 1998. ISSN 01621459. Disponível em: ⟨http://www.jstor.org/stable/2669632⟩.

Freedman, D. A. *Statistical models: theory and practice*. [S.l.]: cambridge university press, 2009.

Hastie, T. J. Generalized additive models. In: *Statistical models in S*. [S.l.]: Routledge, 2017. p. 249–307.

Hofner, B.; Mayr, A.; Schmid, M. gamboostLSS: An R package for model building and variable selection in the GAMLSS framework. *Journal of Statistical Software*, v. 74, n. 1, p. 1–31, 2016.

Johnson, N. L.; Kemp, A. W.; Kotz, S. *Univariate discrete distributions*. [S.l.]: John Wiley & Sons, 2005.

Johnson, N. L.; Kotz, S.; Balakrishnan, N. *Continuous univariate distributions*. [S.l.]: John wiley & sons, 1994. v. 1.

Johnson, N. L.; Kotz, S.; Balakrishnan, N. *Continuous univariate distributions*. [S.l.]: John wiley & sons, 1995. v. 2.

Kutner, M. H. et al. *Applied linear statistical models*. [S.l.]: McGraw-hill, 2005.

Marra, G.; Radice, R. Gjrm: Generalised joint regression modelling. *R Package*, p. R package version 0.2–6.4, 2023.

Nelder, J. A.; Wedderburn, R. W. M. Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, [Royal Statistical Society, Wiley], v. 135, n. 3, p. 370–384, 1972. ISSN 00359238. Disponível em: ⟨http://www.jstor.org/stable/2344614⟩.

Paula, G. A. *Modelos de regressão: com apoio computacional*. [S.l.]: IME-USP São Paulo, 2023.

Ramanathan, R. *Statistical Methods in Econometrics*. [S.l.]: Emerald Publishing, 1993.

Rigby, R. A.; Stasinopoulos, D. M. Smooth centile curves for skew and kurtotic data modelled using the box–cox power exponential distribution. *Statistics in Medicine*, v. 23, n. 19, p. 3053–3076, 2004. Disponível em: ⟨https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.1861⟩.

Rigby, R. A.; Stasinopoulos, D. M. Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society Series C: Applied Statistics*, Oxford University Press, v. 54, n. 3, p. 507–554, 2005.

Rigby, R. A. et al. *Distributions for modeling location, scale, and shape: Using GAMLSS in R*. [S.l.]: CRC press, 2019.

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

Stasinopoulos, D. M.; Rigby, R. A.; Fahrmeir, L. Modelling rental guide data using mean and dispersion additive models. *Journal of the Royal Statistical Society. Series D (The Statistician)*, [Royal Statistical Society, Wiley], v. 49, n. 4, p. 479–493, 2000. ISSN 00390526, 14679884. Disponível em: ⟨http://www.jstor.org/stable/2681031⟩.

Stasinopoulos, M. et al. *Generalized Additive Models for Location, Scale and Shape: A Distributional Regression Approach, with Applications*. [S.l.]: Cambridge University Press, 2024.

Stasinopoulos, M. et al. Package 'gamlss'. 2023.

Stasinopoulos, M. D. et al. *Flexible regression and smoothing: using GAMLSS in R*. [S.l.]: CRC Press, 2017.

Umlauf, N.; Klein, N.; Zeileis, A. Bamlss: Bayesian additive models for location, scale, and shape (and beyond). *Journal of Computational and Graphical Statistics*, Taylor & Francis, v. 27, n. 3, p. 612–627, 2018.

Yee, T. W. *VGAM: Vector Generalized Linear and Additive Models*. [S.l.], 2024. R package version 1.1-11. Disponível em: ⟨https://CRAN.R-project.org/package=VGAM⟩.