

Universidade de Brasília
Departamento de Estatística

Estudo sobre Modelos de Regressão para Dados de Contagem

Murilo Pimentel Massaro

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
2025

Murilo Pimentel Massaro

Estudo sobre Modelos de Regressão para Dados de Contagem

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
2025

Agradecimentos

- Aos meus pais, Marlon e Márcia, pelo apoio incondicional durante todo meu percurso.
- Aos amigos que fiz durante a graduação, que tornaram esse período maravilhoso.
- Aos meus professores, pela dedicação e compartilhamento de conhecimento.
- À minha orientadora, Terezinha, pela paciência, parceria e suporte durante todo o desenvolvimento do trabalho.

Resumo

O tratamento estatístico de variáveis de contagem exige modelos capazes de lidar com a natureza discreta desses dados e com padrões frequentemente observados, como sobredispersão, subdispersão e excesso de zeros. Diversas distribuições são apresentadas para acomodar essas características, assim como métodos de regressão que permitem relacionar a variável resposta a covariáveis de interesse. Este trabalho apresenta uma revisão das principais abordagens utilizadas na modelagem de contagens, comentando sobre o uso dos modelos lineares generalizados (MLGs) e análise diagnóstica desses modelos. São abordados também os modelos aditivos generalizados para localização, escala e forma (GAMLSS), que permitem modelar múltiplos parâmetros da distribuição. Essa flexibilidade torna essa classe de modelos particularmente adequada para cenários mais complexos, nos quais a média, a variabilidade ou mesmo características como assimetria e curtose podem depender de covariáveis. Além disso, técnicas de inferência, critérios para comparação entre modelos e métodos de avaliação da qualidade do ajuste também são descritos, com destaque para os resíduos quantílicos, amplamente utilizados no diagnóstico dos GAMLSSs. Para ilustrar a aplicação prática dessas metodologias, dois conjuntos de dados reais são analisados, evidenciando como diferentes distribuições podem capturar distintos padrões presentes nos dados. O estudo busca oferecer uma visão abrangente das possibilidades de modelagem para dados de contagem, destacando tanto suas potencialidades quanto suas limitações.

Palavras-chave: Análise diagnóstica; GAMLSS; Inflação de zeros; Modelos de contagem; Resíduos quantílicos; Sobredispersão.

Lista de Tabelas

1 Funções de ligação disponíveis no pacote `gamlss`, de acordo com o domínio dos parâmetros da distribuição. 17

2 Distribuições discretas de contagem, suas abreviações e parâmetros. 17

3 Interpretação do ajuste de um modelo baseado em diferentes formas do *worm plot*. 37

4 Valores do critério de informação AIC para diferentes distribuições discretas de contagem ajustadas marginalmente para o número de pólipos do paciente. 42

5 Estimativas, erros-padrão, estatísticas z e p-valores dos coeficientes sob o Modelo 3 da Aplicação 1. 47

6 Valores do critério de informação AIC para diferentes distribuições discretas de contagem ajustadas marginalmente para o número de satélites de caranguejo-ferradura. 50

7 Estimativas, erros-padrão, estatísticas z e p-valores dos coeficientes sob o Modelo 5. 59

Lista de Figuras

1 Diagrama de classificação das distribuições de contagem. 32

2 Gráfico de probabilidade normal com envelope simulado para os resíduos desvio do modelo de regressão Poisson ajustado. 36

3 *Worm plot* do modelo de exemplo 37

4 Gráficos de diagnóstico para os resíduos quantílicos sob o modelo de exemplo: (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. 38

5 (a) Gráfico de dispersão entre o número de pólipos e idade em anos do paciente; (b) *boxplot* do número de pólipos e tipo de tratamento utilizado. 41

6 Gráficos de diagnóstico para os resíduos quantílicos sob o Modelo 1 da Aplicação 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. 43

7 *Worm plot* dos resíduos quantílicos sob o Modelo 1 da Aplicação 1. 43

8 Gráficos de diagnóstico para os resíduos quantílicos sob o Modelo 2 da Aplicação 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. 44

9 *Worm plot* dos resíduos quantílicos sob o Modelo 2 da Aplicação 1. 45

10 Gráficos de diagnóstico para os resíduos quantílicos sob o Modelo 3 da Aplicação 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. 46

11 *Worm plot* dos resíduos quantílicos sob o Modelo 3 da Aplicação 1. 47

12 Gráfico de dispersão entre o número de satélites e (a) peso do caranguejo em quilogramas e *boxplot* do número de satélites pela (b) cor da carapaça. 49

13 Gráficos de diagnóstico para os resíduos quantílicos sob o Modelo 1 da Aplicação 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. 51

14 *Worm plot* dos resíduos quantílicos sob o Modelo 1 da Aplicação 2. 51

15 Gráficos de diagnóstico para os resíduos quantílicos sob o Modelo 2 da Aplicação 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. 52

16 *Worm plot* dos resíduos quantílicos sob o Modelo 2 da Aplicação 2. 53

17 Gráficos de diagnóstico para os resíduos quantílicos sob o Modelo 3 da Aplicação 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. 54

18 *Worm plot* dos resíduos quantílicos sob o Modelo 3 da Aplicação 2. 55

19 Gráficos de diagnóstico para os resíduos quantílicos sob o Modelo 4 da Aplicação 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. 56

20 *Worm plot* dos resíduos quantílicos sob o Modelo 4 da Aplicação 2. 56

21 Gráficos de diagnóstico para os resíduos quantílicos sob o Modelo 5 da Aplicação 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. 58

22 *Worm plot* dos resíduos quantílicos sob o Modelo 5 da Aplicação 2. 58

Sumário

1 Introdução 8

2 Referencial Teórico 10

2.1 Modelo de regressão linear normal 10

2.2 Modelos lineares generalizados 12

2.3 Modelos aditivos generalizados para localização, escala e forma 14

2.4 Distribuições inflacionadas e ajustadas em zero 18

2.5 Distribuições discretas de contagem 19

2.5.1 Poisson 19

2.5.2 Poisson Ajustada em Zero 19

2.5.3 Poisson Inflacionada em Zero 20

2.5.4 Geométrica 21

2.5.5 Binomial Negativa 21

2.5.6 Binomial Negativa Ajustada em Zero 22

2.5.7 Binomial Negativa Inflacionada em Zero 22

2.5.8 Poisson Generalizada 22

2.5.9 Delaporte 23

2.5.10 Família Binomial Negativa 23

2.5.11 Poisson Dupla 24

2.5.12 Burr XII Discreta 24

2.5.13 Poisson-Gaussiana Inversa 25

2.5.14 Poisson-Gaussiana Inversa Ajustada em Zero 25

2.5.15 Poisson-Gaussiana Inversa Inflacionada em Zero 26

2.5.16 Sichel 26

2.5.17 Sichel Ajustada em Zero 27

2.5.18 Sichel Inflacionada em Zero 27

2.5.19 Beta-Binomial Negativa 27

2.5.20 Beta-Binomial Negativa Ajustada em Zero 28
2.5.21 Beta-Binomial Negativa Inflacionada em Zero 28
2.5.22 Yule 29
2.5.23 Waring 29
2.5.24 Logarítmica 30
2.5.25 Logarítmica Ajustada em Zero 30
2.5.26 Zipf 31
2.5.27 Zipf Ajustada em Zero 31
2.5.28 Síntese das distribuições de contagem 32
2.6 Métodos de análise diagnóstica e avaliação da qualidade do ajuste 33
2.6.1 Medida de alavancagem 33
2.6.2 Resíduos 33
2.6.3 Técnicas gráficas 35
3 Aplicações 40
3.1 Aplicação 1: Número de pólipos em função da idade e do tipo de tratamento 40
3.2 Aplicação 2: Número de satélites em função das características das fêmeas de caranguejo-ferradura 49
4 Conclusão 62
Referências 64

1 Introdução

A regressão linear normal é um dos métodos mais comumente utilizados na modelagem de relações entre variáveis (KUTNER et al., 2005). Entretanto, a aplicação dessa técnica a dados de contagem não é a mais adequada e pode gerar resultados imprecisos. Dados de contagem são discretos e assumem apenas valores não negativos, enquanto a regressão linear normal supõe que a variável resposta possa variar em todo o conjunto dos reais. Além disso, o modelo linear normal pressupõe simetria e homoscedasticidade, o que muitas vezes não ocorre com dados de contagem. Dessa forma, a utilização da regressão linear em casos de contagem pode levar a inferências distorcidas e interpretações incondizentes.

Para acomodar uma maior variedade de estruturas de dados, foram desenvolvidos os modelos lineares generalizados (MLGs) e suas extensões. Essa classe de modelos representa uma generalização do modelo linear clássico, permitindo que a variável resposta siga qualquer distribuição pertencente à família exponencial uniparamétrica (PAULA, 2025). Os MLGs flexibilizam as suposições sobre a distribuição dos dados e são especialmente úteis quando a variável resposta não apresenta distribuição normal, como é comum em dados de contagem. Nesse contexto, destacam-se as distribuições Poisson e binomial negativa, amplamente utilizadas no ajuste de modelos para esse tipo de dado. Apesar da flexibilidade oferecida pelos MLGs ao permitir diferentes distribuições para a variável resposta e a introdução da função de ligação, essa classe de modelos ainda apresenta algumas limitações importantes. Entre elas, destaca-se a restrição à modelagem de apenas um parâmetro da distribuição, geralmente o parâmetro de localização, o que pode ser insuficiente em situações onde a variabilidade ou a assimetria dos dados também dependem de covariáveis. Além disso, os MLGs pressupõem uma estrutura linear no preditor, o que pode não capturar adequadamente relações mais complexas entre as variáveis explicativas e a resposta. Essas limitações motivaram o desenvolvimento de abordagens mais gerais capazes de lidar com estruturas de dados mais heterogêneas e com maior complexidade.

Para contornar essas dificuldades, modelos mais flexíveis são utilizados, no caso deste trabalho serão discutidos os modelos aditivos generalizados para localização, escala e forma (*Generalized Additive Models for Location, Scale and Shape*; GAMLSS) (RIGBY et al., 2019). Os GAMLSSs permitem que não apenas a média, mas também outros parâmetros da distribuição, como dispersão, curtose e assimetria, sejam modelados em função de covariáveis, oferecendo maior capacidade de ajuste e interpretação em cenários complexos. Tais como as modelagens para dados de contagem pelos métodos citados

possuem diversas aplicações como na modelagem de acidentes de trânsito (COSTA et al., 2012), na epidemiologia e saúde pública (COSTA; SILVEIRA; DONALISIO, 2016), na biologia (PEDRO; SILVEIRA; TASSINARI, 2014), entre outros campos de estudo.

Este trabalho visa realizar uma revisão sobre os principais modelos de regressão para dados de contagem, discutir brevemente suas características, suposições, limitações e análise diagnóstica. Alguns dos modelos de regressão que o trabalho se propõe a estudar são, entre os MLGs, os modelos de Poisson, Poisson truncado e extensões dos MLGs, como o modelo binomial negativo, considerando os GAMLSSs, o modelo de Poisson dupla e modelos ajustados e inflacionados em zero. O restante desse relatório está organizado como se segue. No Capítulo 2 é feita a revisão teórica que fundamenta os conceitos e métodos estatísticos utilizados ao longo do trabalho, incluindo a descrição de diversas distribuições de contagem empregadas na modelagem. No Capítulo 3, são apresentas duas aplicações a conjuntos de dados reais que exemplificam o uso das metodologias apresentadas e demonstram sua utilização prática. Por fim, no Capítulo 4, são apresentadas as considerações finais do trabalho. Cabe citar, ainda, que os ajustes de regressão do trabalho foram realizados através do software `R`, um software de domínio público disponível em https://cran.r-project.org/. Os códigos em `R` empregados nas aplicações desenvolvidas neste trabalho estão disponíveis em um repositório público no *GitHub*, no endereço: https://github.com/MPMassaro/TCC-Regressao-de-Contagem.

2 Referencial Teórico

O relatório tem como objetivo fazer uma revisão sobre modelos de regressão para dados de contagem. Nesse sentido, para se ter melhor entendimento do conteúdo apresentado neste relatório, será feita uma breve revisão dos principais pontos abordando modelos lineares normais, modelos lineares generalizados (MLGs) e modelos aditivos generalizados de localização, escala e forma (GAMLSSs). Além disso, também será realizada uma revisão das distribuições adequadas para dados de contagem e dos métodos para análise diagnóstica e de qualidade do ajuste dos modelos em estudo.

2.1 Modelo de regressão linear normal

A regressão linear é uma técnica estatística utilizada para modelar a média de uma variável resposta Y em função de uma ou mais variáveis explicativas $X_1, X_2, \ldots, X_k$. Quando é ajustado por apenas uma variável explicativa, o modelo é denominado regressão linear simples, e na presença de múltiplas variáveis explicativas, denomina-se regressão linear múltipla (KUTNER et al., 2005).

O modelo de regressão linear múltipla é representado pela seguinte equação:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i, \quad i = 1, 2, \ldots, n,$$

ou, de forma matricial

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

em que

- $\mathbf{Y}$ é um vetor $n \times 1$ contendo as n observações da variável resposta;
- $\mathbf{X}$ é uma matriz não aleatória de dimensão $n \times (k+1)$, em que a primeira coluna é formada por 1's corresponde ao intercepto e as demais k colunas são correspondentes às observações das variáveis explicativas, além disso, essa matriz deve ter posto completo;
- $\boldsymbol{\beta}$ é o vetor de parâmetros desconhecidos (coeficientes) do modelo, com dimensão $(k+1) \times 1$;
- $\boldsymbol{\varepsilon}$ é o vetor de erros aleatórios, assumindo-se que seguem independentemente distribuição normal com média zero e variância σ^2 constante.

A estimativa dos coeficientes $\boldsymbol{\beta}$ é obtida pelo método dos mínimos quadrados, que minimiza a soma dos quadrados dos erros definida por (KUTNER et al., 2005)

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}.$$

A partir da estimação dos coeficientes da regressão, é possível realizar previsões, avaliar a significância dos preditores e verificar a adequação do modelo por meio de análise diagnóstica.

Além disso, o estimador do parâmetro da variância dos erros σ^2 não é obtido diretamente a partir do método de mínimos quadrados. Porém, pode-se obter um estimador de forma indireta através de

$$\hat{\sigma}^2 = \frac{1}{n-k-1} \sum_{i=1}^{n} \hat{\varepsilon}_i^2,$$

em que $\hat{\varepsilon}_i = Y_i - \hat{Y}_i$, $i = 1, 2, ..., n$. Esse estimador indireto pode ser usado para encontrar o estimador de máxima verossimilhança de σ^2, por meio de

$$\hat{\sigma}^2_{MV} = \frac{(n-k-1)\hat{\sigma}^2}{n}.$$

No modelo linear normal, os testes de hipótese e os intervalos de confiança para os coeficientes são construídos a partir da suposição de que os erros seguem distribuição normal com variância constante. Dessa forma, as estimativas dos parâmetros possuem distribuição t de Student, permitindo realizar teste de hipótese acerca da significância do coeficiente no modelo, ou seja, $H_0 : \beta_j = 0$ contra $H_1 : \beta_j \neq 0$, por meio da estatística $t = \hat{\beta}_j / \sqrt{\hat{\sigma}^2_{MV} \ \mathrm{a}_{(\mathrm{j+1})(\mathrm{j+1})}}$, em que $\mathrm{a}_{(\mathrm{j+1})(\mathrm{j+1})}$ é o $(j+1)$-ésimo elemento da diagonal principal da matriz $(\mathbf{X}^\top \mathbf{X})^{-1}$, com $j = 0, 1, 2, ..., k$. De forma análoga, os intervalos de confiança são obtidos na forma

$$\hat{\beta}_j \pm t_{\ n-k-1,\ 1-\alpha/2} \sqrt{\hat{\sigma}^2_{MV} \ \mathrm{a}_{(\mathrm{j+1})(\mathrm{j+1})}},$$

em que $t_{\ n-k-1,\ 1-\alpha/2}$ é o quantil da distribuição t com $n-k-1$ graus de liberdade.

O modelo também está sujeito a alguns pressupostos que devem ser cumpridos para que as estimativas obtidas e outras inferências sejam confiáveis. Algumas dessas pressuposições podem limitar a flexibilidade do ajuste desses modelos aos dados. Nesse sentido, podem-se citar as suposições de normalidade e homoscedasticidade dos erros, que muitas vezes não são atendidas. Além disso, para dados de contagem, na maioria dos casos, a relação entre as variáveis resposta e explicativa não é linear, o que representa um grande problema para o modelo linear normal (PAULA, 2025). Assim, os modelos

abordados a seguir, por serem mais flexíveis, acomodam melhor os dados de contagem gerando ajustes mais adequados e resultados mais confiáveis.

2.2 Modelos lineares generalizados

Analogamente aos modelos de regressão linear, os modelos lineares generalizados buscam compreender o comportamento de uma variável resposta Y em relação a uma ou mais variáveis explicativas. No entanto, a variável resposta pode seguir outras distribuições além da distribuição normal, inclusive distribuições discretas mais adequadas para dados de contagem e a relação entre média da resposta e preditor linear é flexibilizada através da função de ligação. A variável resposta deve seguir alguma distribuição que pertença à família exponencial uniparamétrica (CORDEIRO; DEMÉTRIO, 2008). A função densidade de probabilidade da variável resposta Y_i deve ser expressa por

$$f(y_i;\theta_i,\phi) = \exp\left\{\frac{1}{\phi}\left[y_i\theta_i - b(\theta_i)\right] + c(y_i,\phi)\right\}, \tag{2.2.1}$$

em que

- y_i é a i-ésima observação da variável resposta;
- θ_i é o parâmetro canônico da distribuição tal que $\theta_i = \theta_i(\mu_i)$ função de $\mu_i = \mathrm{E}(Y_i)$;
- $b(\theta_i)$ é uma função de θ_i tal que $b'(\theta_i) = db(\theta_i)/d\theta_i = \mathrm{E}(Y_i)$;
- $c(y_i,\phi)$ é uma função de y_i e ϕ;
- ϕ é o parâmetro de dispersão da distribuição, invariante para cada observação i, tal que $\mathrm{Var}(Y_i) = \phi b''(\theta_i) = d^2b(\theta_i)/d\theta_i^2 = \phi V(\mu_i)$, sendo $V(\mu_i)$ a função de variância.

Os modelos lineares generalizados são definidos por (3.2.1) e pela parte sistemática

$$g(\mu_i) = \eta_i, \tag{2.2.2}$$

em que $\eta_i = \mathbf{X}_i^\top\boldsymbol{\beta}$ é o preditor linear, $\boldsymbol{\beta} = (\beta_0,\ldots,\beta_k)^\top$, $k+1 < n$, é um vetor de parâmetros desconhecidos a serem estimados, $\mathbf{X}_i = (1, x_{i1},\ldots,x_{ik})^\top$ representa os valores de variáveis explicativas para a i-ésima observação, e $g(\cdot)$ é uma função monótona e diferenciável, denominada função de ligação. Essa função tem como base a ideia de mapear o espaço em que está o parâmetro de localização para a reta real, o que garante que os parâmetro de regressão sejam estimados sem restrições.

A estimação dos parâmetros é feita pelo método da máxima verossimilhança, assumindo que a distribuição da variável resposta pertence à família exponencial uniparamétrica. Assim, sendo $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top \ \phi)^\top$ o logaritmo da função de verossimilhança para a amostra é dado por

$$\ell(\boldsymbol{\theta}) = \phi \sum_{i=1}^{n} [y_i\theta_i - b(\theta_i)] + \sum_{i=1}^{n} c(y_i; \phi).$$

Para encontrar o valor de $\boldsymbol{\theta}$ que maximiza $\ell(\boldsymbol{\theta})$, deriva-se $\ell(\boldsymbol{\theta})$ com relação à $\boldsymbol{\theta}$ e, então, iguala-se ao vetor de zeros de dimensão $(k+1)$. O vetor escore para $\boldsymbol{\beta}$ fica dado por

$$U_{\boldsymbol{\beta}}(\boldsymbol{\theta}) = \frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}} = \phi X^\top W^{1/2} V^{-1/2} (\boldsymbol{y} - \boldsymbol{\mu}),$$

em que V é a matriz diagonal com as variâncias de cada observação expressada como $V = \text{diag}\{\text{Var}(Y_1), \text{Var}(Y_2), ..., \text{Var}(Y_n)\} = \text{diag}\{\phi V(\mu_1), ..., \phi V(\mu_n)\}$ e W é a matriz de pesos diagonal com os pesos denotada por $W = \text{diag}\{\omega_1, ..., \omega_n\}$, com $\omega_i = \frac{(d\mu_i/d\eta_i)^2}{V_i}$,

$$X = \begin{bmatrix} \boldsymbol{x}_1^T \\ \boldsymbol{x}_2^T \\ \vdots \\ \boldsymbol{x}_n^T \end{bmatrix}, \quad \boldsymbol{x}_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{bmatrix}, \quad \boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix}$$

enquanto o vetor escore para ϕ fica dado por

$$U_\phi(\boldsymbol{\theta}) = \frac{\partial \ell(\boldsymbol{\theta})}{\partial \phi} = \sum_{i=1}^{n} [y_i\theta_i - b(\theta_i)] + \sum_{i=1}^{n} c'(y_i; \phi),$$

em que $c'(y_i; \phi) = dc(y_i; \phi)/d\phi$. Essas quantidades são fundamentais para métodos iterativos de estimação, como o algoritmo escore de Fisher, que é expressado por

$$\beta^{(m+1)} = \beta^{(m)} + \left\{K_{\beta\beta}^{-1}\right\}^{(m)} U_\beta^{(m)}, \quad m = 0, 1, 2, ...,$$

em que

$$U_\beta^{(m)} = \phi X^\top \left\{W^{1/2}\right\}^{(m)} \left\{V^{-1/2}\right\}^{(m)} \left(y - \mu^{(m)}\right),$$

e a matriz de informação de Fisher para $\boldsymbol{\beta}$ é

$$K_{\beta\beta} = -\,\text{E}\left\{\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top}\right\} = \phi X^\top W X.$$

A maximização da função de verossimilhança é realizada, então, por meio do algoritmo de Fisher, um procedimento iterativo baseado em mínimos quadrados reponderados que atualiza sucessivamente os estimadores até atingir a convergência (PAULA, 2025).

A flexibilidade dos MLGs e suas extensões os tornam adequados para diversas situações práticas, uma delas sendo modelos específicos para dados de contagem, como os modelos Poisson, Poisson Truncado, que são MLGs propriamente ditos, e Binomial Negativo, que é uma extensão dos MLGs. Ainda assim, os MLGs, apesar de bastante flexíveis e se apresentarem como ferramentas muito úteis para a análise de dados de contagem, estes possuem algumas limitações, sendo a principal delas a necessidade de a variável resposta seguir distribuição que faça parte da família exponencial uniparamétrica. Dessa forma, na próxima seção será discutida a classe dos modelos aditivos generalizados para localização, escala e forma que ampliam ainda mais a variedade de modelos específicos para dados de contagem.

2.3 Modelos aditivos generalizados para localização, escala e forma

Os modelos aditivos generalizados para localização, escala e forma (*Generalized Additive Models for Location, Scale and Shape*; GAMLSS), propostos por Rigby e Stasinopoulos (2005), constituem uma generalização dos modelos lineares generalizados e dos modelos aditivos generalizados (GAMs), permitindo maior flexibilidade na modelagem estatística. Diferentemente dos MLGs tradicionais, nos quais apenas o parâmetro de localização é modelado como função das covariáveis, os GAMLSSs permitem que vários parâmetros da distribuição, como de localização, dispersão, curtose e assimetria, sejam modelados.

Formalmente, sob os GAMLSSs assume-se que a variável resposta Y_i, com $i = 1, 2, ..., n$, segue uma distribuição $\mathrm{D}(\mu_i, \sigma_i, \nu_i, \tau_i)$ de até quatro parâmetros, os quais podem depender de funções das covariáveis observadas. Cada parâmetro da distribuição é relacionado ao respectivo preditor por meio de uma função de ligação adequada (RIGBY et al., 2019):

$$
\begin{aligned}
g_1(\boldsymbol{\mu}) &= \boldsymbol{\eta}_1 = \mathbf{X}_1\boldsymbol{\beta}_1 + s_{11}(\mathbf{x}_{11}) + \cdots + s_{1J_1}(\mathbf{x}_{1J_1}),\\
g_2(\boldsymbol{\sigma}) &= \boldsymbol{\eta}_2 = \mathbf{X}_2\boldsymbol{\beta}_2 + s_{21}(\mathbf{x}_{21}) + \cdots + s_{2J_2}(\mathbf{x}_{2J_2}),\\
g_3(\boldsymbol{\nu}) &= \boldsymbol{\eta}_3 = \mathbf{X}_3\boldsymbol{\beta}_3 + s_{31}(\mathbf{x}_{31}) + \cdots + s_{3J_3}(\mathbf{x}_{3J_3}),\\
g_4(\boldsymbol{\tau}) &= \boldsymbol{\eta}_4 = \mathbf{X}_4\boldsymbol{\beta}_4 + s_{41}(\mathbf{x}_{41}) + \cdots + s_{4J_4}(\mathbf{x}_{4J_4}).
\end{aligned}
$$

em que

- η_k é o k-ésimo preditor relacionado ao k-ésimo parâmetro da distribuição, $k =$

$1, 2, 3, 4$;

- $\mathbf{X}_k = [\mathbf{x}_{1k}\ \mathbf{x}_{2k}\ ...\ \mathbf{x}_{p_k k}]$ é uma matriz $n \times p_k$ contendo os n valores fixados das p_k das covariáveis do k-ésimo submodelo;

- $\mathbf{x}_{jk} = (x_{1jk}\ x_{2jk}\ ...\ x_{njk})^\top \in \mathbb{R}^n$, $j = 1, ..., p_k$, é o vetor que contém os n valores fixados da j-ésima covariável associado ao k-ésimo submodelo;

- β_k é o vetor de coeficientes da regressão correspondente ao p-ésimo submodelo;

- $s_{kj}(\mathbf{x}_{kj})$ com $j = 1, ..., J_k$ podem ser lineares ou não lineares, como funções de suavização, de covariáveis.

A possibilidade de os preditores η_k serem expressos como somas de termos lineares e não lineares confere aos GAMLSS grande capacidade de adaptação a diferentes formatos de dados.

Nos GAMLSSs, a estimação dos parâmetros é realizada por meio da máxima verossimilhança penalizada, permitindo o ajuste simultâneo de vários parâmetros da distribuição, cada um modelado como função de covariáveis. Para os GAMLSS paramétricos, os parâmetros são estimados via máxima verossimilhança. O logaritmo da função de verossimilhança, considerando a independência de $Y_1, Y_2, \ldots, Y_n$, é (RIGBY et al., 2019)

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

em que $\boldsymbol{\theta} = (\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top, \boldsymbol{\beta}_3^\top, \boldsymbol{\beta}_4^\top)^\top$, $f(y_i; \mu_i, \sigma_i, \nu_i, \tau_i)$ é a função densidade da i-ésima observação da resposta com distribuição $\mathrm{D}(\mu_i, \sigma_i, \nu_i, \tau_i)$. A estimação é realizada por meio de algoritmos iterativos, como o algoritmo de Cole e Green (CG) ou o algoritmo de Rigby e Stasinopoulos (RS) (STASINOPOULOS et al., 2017). Os parâmetros de suavização são escolhidos automaticamente com base em critérios como o Critério de Informação de Akaike (AIC) ou o Critério de Informação de Akaike Generalizado (GAIC), que se trata de uma generalização do AIC para comparação de modelos com *overfit*, ou ainda por métodos baseados em perfil de verossimilhança marginal aproximada. Essa abordagem garante grande flexibilidade ao modelo, sendo adequada para dados com estruturas complexas de dispersão, assimetria ou cauda.

A estimação por máxima verossimilhança fornece os valores pontuais dos coeficientes do modelo, mas a inferência estatística requer a quantificação da incerteza associada a essas estimativas. Essa quantificação é fundamental para avaliar a significância dos parâmetros e construir intervalos de confiança. Sob condições de regularidade, os

estimadores de máxima verossimilhança possuem propriedades assintóticas ideais: são consistentes, invariantes, eficientes e distribuem-se normalmente para grandes amostras. Considerando o vetor $\boldsymbol{\theta}$, que nesse contexto representa a coleção de todos os coeficientes lineares dos submodelos correspondentes aos parâmetros da distribuição $(\mu_i, \sigma_i, \nu_i, \tau_i)$, como, $\boldsymbol{\theta} = \left(\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top, \boldsymbol{\beta}_3^\top, \boldsymbol{\beta}_4^\top\right)^\top$. Assintoticamente, a distribuição do estimador de máxima verossimilhança $\hat{\boldsymbol{\theta}}$ é dada por

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

em que $\boldsymbol{\theta}_T$ denota o verdadeiro valor do parâmetro e $\boldsymbol{i}(\boldsymbol{\theta}_T)$ é a matriz de informação de Fisher esperada, dada por

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

Como nem sempre é possível derivar a informação esperada $\boldsymbol{i}(\boldsymbol{\theta}_T)$ analiticamente, a informação de Fisher observada $\boldsymbol{I}(\boldsymbol{\theta}_T)$, definida como

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

é calculada diretamente a partir do negativo da matriz Hessiana da função do logaritmo da função verossimilhança. Como $\boldsymbol{\theta}_T$ é desconhecido, ambas as matrizes são avaliadas no estimador $\hat{\boldsymbol{\theta}}$, resultando em $\boldsymbol{i}(\hat{\boldsymbol{\theta}})$ e $\boldsymbol{I}(\hat{\boldsymbol{\theta}})$. Em modelos GAMLSS paramétricos, a variância assintótica de $\hat{\boldsymbol{\theta}}$ é usualmente aproximada pela inversa da matriz de informação observada

$$\hat{\boldsymbol{\theta}} \sim \mathcal{N}\Big(\boldsymbol{\theta}_T,\ \boldsymbol{I}(\hat{\boldsymbol{\theta}})^{-1}\Big).$$

Assim, este resultado estabelece a base para os principais instrumentos de inferência, como predição intervalar, construção de intervalos de confiança de Wald e testes de hipóteses sobre a significância dos covariáveis.

A Tabela 1 apresenta as função de ligação para os submodelos dos parâmetros disponíveis no pacote `gamlss`, de acordo com o domínio dos parâmetros da distribuição. Nos GAMLSS, as funções de ligação conectam os preditores lineares aos parâmetros da distribuição, garantindo que as estimativas respeitem o domínio de cada parâmetro. As funções de ligação usuais, como a logarítmica para parâmetros positivos e a logit para parâmetros no intervalo unitário, são vantajosas por facilitarem a interpretação dos coeficientes. No caso da ligação logarítmica, os efeitos são interpretados em termos da mudança relativa, enquanto a logit permite interpretar os efeitos em termos da razão de chances ou ainda da própria probabilidade.

Tabela 1: Funções de ligação disponíveis no pacote `gamlss`, de acordo com o domínio dos parâmetros da distribuição.

Domínio do parâmetro	Função de ligação	Fórmula para $g(\theta)$
$-\infty$ a ∞	identidade	θ
0 a ∞	log	$\log(\theta)$
	sqrt	$\sqrt{\theta}$
	inversa	$1/\theta$
	$1/\mu^2$	$1/\theta^2$
	μ^2	θ^2
0 a 1	logit	$\log\left(\frac{\theta}{1-\theta}\right)$
	probit	$\Phi^{-1}(\theta)$
	cauchit	$\tan(\pi(\theta - 0{,}5))$
	cloglog	$\log(-\log(1-\theta))$

Uma das principais vantagens dos GAMLSSs em relação aos MLGs é o fato de não se limitar às distribuições da família exponencial uniparamétrica, o que amplia significativamente o conjunto de distribuições disponíveis utilizáveis para modelagem. Atualmente, há dezenas de distribuições implementadas no software `R` no pacote `gamlss` (STASINOPOULOS et al., 2024), elaborado pelos próprios autores. Vale citar, ainda, que toda a aplicação computacional feita no trabalho referente ao GAMLSS foi realizada por meio do pacote `gamlss` pelo software `R`. A Tabela 2 apresenta as distribuições discretas de contagem que estão presentes no pacote `gamlss` do `R`, suas abreviações e parâmetros.

Tabela 2: Distribuições discretas de contagem, suas abreviações e parâmetros.

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)
Poisson Dupla	DPO(μ,σ)	Logarítmica Ajustada em Zero	ZALG(μ)
Poisson Generalizada	GPO(μ,σ)	Binomial Negativa Ajustada em Zero	ZANBI(μ,σ)
Geométrica	GEOM(μ)	Poisson-Gaussiana Inversa Ajustada em Zero	ZAPIG(μ,σ)
Geométrica Original	GEOMO(μ)	Sichel Ajustada em Zero	ZASICHEL(μ,σ,ν)
Logarítmica	LG(μ)	Poisson Ajustada em Zero	ZAP(μ)
Binomial Negativa	NBI(μ,σ)	Zipf Ajustada em Zero	ZAZIPF(μ)
Binomial Negativa Reparametrizada	NBII(μ,σ)	Beta-Binomial Negativa Inflacionada em Zero	ZIBNB(μ,σ,ν,τ)
Família Binomial Negativa	NBF(μ,ν)	Binomial Negativa Inflacionada em Zero	ZINBI(μ,σ,τ)
Poisson	PO(μ)	Família Binomial Negativa Inflacionada em Zero	ZINBF(μ,ν,τ)
Poisson-Gaussiana Inversa	PIG(μ,σ)	Poisson Inflacionada em Zero	ZIP(μ,τ)
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,τ)
Sichel Reparametrizada	SICHEL(μ,σ,ν)	Sichel Inflacionada em Zero	ZISICHEL(μ,σ,ν,τ)
Waring	WARING(μ,σ)		

Para dados de contagem, os GAMLSSs oferecem ferramentas importantes ao permitir modelar não apenas a média, mas também a curtose, assimetria e dispersão. Nessa

perspectiva, muitos modelos disponíveis na classe GAMLSS superam uma das principais limitações do MLG Poisson, que assume equidispersão, isto é, $\mathrm{Var}(Y_i) = \mathrm{E}(Y_i)$. Diversos modelos do GAMLSS oferecem distribuições que acomodam adequadamente dados cuja variável resposta apresenta sobredispersão, $\mathrm{Var}(Y_i) > \mathrm{E}(Y_i)$, ou subdispersão, $\mathrm{Var}(Y_i) < \mathrm{E}(Y_i)$. Além das tradicionais distribuições de Poisson e binomial negativa, estão disponíveis também distribuições inflacionadas em zero para dados com muitos valores nulos.

2.4 Distribuições inflacionadas e ajustadas em zero

Em muitas aplicações práticas envolvendo dados de contagem, observa-se uma frequência de valores nulos significativamente maior ou menor do que a prevista por uma distribuição discreta tradicional adequada para dados de contagem (RIGBY et al., 2019). Para lidar com essa característica dos dados, foram elaboradas distribuições inflacionadas em zero e distribuições ajustadas em zero.

Distribuições inflacionadas em zero são modelos de mistura que combinam duas componentes, uma massa de probabilidade adicional concentrada no zero e uma distribuição de contagem tradicional para os demais valores. Seja Y uma variável aleatória discreta de contagem, então a distribuição inflacionada em zero é dada por (RIGBY et al., 2019)

$$\mathrm{P}(Y = y) = \begin{cases} p + (1-p)\mathrm{P}(Y_1 = 0), & \text{se } y = 0, \\ (1-p)\mathrm{P}(Y_1 = y), & \text{se } y = 1, 2, 3, \ldots, \end{cases}$$

em que $0 < p < 1$ representa a probabilidade adicional associada ao zero, e Y_1 segue uma dada distribuição de contagem. O uso dessa distribuição inflacionada em zero em modelagem é adequado quando a proporção de zeros observados nos dados é maior do que o previsto pela distribuição original. A média e outros momentos de uma distribuição inflacionada em zero são diretamente influenciados por p. A média, por exemplo, fica dada por

$$\mathrm{E}(Y) = (1-p)\,\mathrm{E}(Y_1).$$

Assim, a presença de inflação em zero reduz a média em relação à distribuição base.

Enquanto a inflação em zero adiciona probabilidade extra no zero, as distribuições ajustadas em zero modificam a probabilidade do zero para que esta possa ser maior ou menor do que a prevista pela distribuição de contagem original. As distribuições ajustadas também são modelos de mistura, mas a probabilidade do zero é diretamente

parametrizada e os valores diferentes de zero seguem a distribuição truncada no zero. Assim, a distribuição ajustada em zero fica dada por (RIGBY et al., 2019)

$$\mathrm{P}(Y=y)=\begin{cases} p, & \text{se } y=0, \\ (1-p)\mathrm{P}(Y_1=y), & \text{se } y=1,2,3,\ldots, \end{cases}$$

em que $0<p<1$ representa a nova probabilidade associada ao zero, e Y_1 segue uma dada distribuição de contagem truncada em zero. Ao contrário das distribuições inflacionadas, as ajustadas em zero não assumem que a probabilidade de ocorrência de zero seja necessariamente maior que a esperada. Distribuições ajustadas em zero são úteis na modelagem de dados de contagem quando existe ausência de contagens nulas ou a frequência é diferente da prevista pela distribuição base.

2.5 Distribuições discretas de contagem

A seguir são detalhadas algumas das distribuições discretas adequadas para dados de contagem que podem ser ajustadas sob os MLGs e os GAMLSSs.

2.5.1 Poisson

A distribuição de Poisson é a distribuição mais amplamente empregada na modelagem de dados de contagem, porém, apresenta algumas limitações. A função de probabilidade é dada por (RIGBY et al., 2019)

$$f(y \mid \mu)=\frac{\mathrm{e}^{-\mu}\mu^{y}}{y!}, \quad y\in\{0,1,2,...\}, \quad \mu>0,$$

em que o parâmetro μ representa tanto a média quanto a variância, além de também ser parâmetro de assimetria e curtose da distribuição. Essa igualdade entre média e variância caracteriza o fenômeno conhecido como equidispersão, o que causa uma das principais limitações dessa distribuição sob os modelos de regressão, além de possuir apenas 1 parâmetro. Nos MLGs e GAMLSSs, o parâmetro μ é usualmente modelado a partir da função de ligação logarítmica.

2.5.2 Poisson Ajustada em Zero

O ajuste em zero da distribuição de Poisson modifica a probabilidade associada ao valor zero, permitindo lidar com situações em que há ausência de zeros ou em que

a frequência desse valor não é adequadamente representada pela distribuição de Poisson tradicional. A função de probabilidade é expressada por (RIGBY et al., 2019)

$$f(y \mid \mu, \sigma) = \begin{cases} \sigma, & \text{se } y = 0, \\ \dfrac{(1-\sigma)}{(1-\mathrm{e}^{-\mu})} \cdot \dfrac{\mathrm{e}^{-\mu}\mu^{y}}{y!}, & \text{se } y = 1, 2, 3, \ldots, \end{cases}$$

em que $\mu > 0$ e $0 < \sigma < 1$. O parâmetro μ representa a média da distribuição Poisson antes do truncamento em zero, e σ a probabilidade de ocorrência de zero. Nesse sentido, a média da distribuição fica dada por $(1-\sigma)\mu/(1-\mathrm{e}^{-1\mu})$, caso ocorra o truncamento em zero, ou seja $\sigma = 0$, a média fica dada por $\mu/(1-\mathrm{e}^{-\mu})$.

2.5.3 Poisson Inflacionada em Zero

A distribuição de Poisson inflacionada em zero é uma das distribuições adequadas para modelagem de dados de contagem com excesso de zeros. A função de probabilidade é dada por (RIGBY et al., 2019)

$$f(y \mid \mu, \sigma) = \begin{cases} \sigma + (1-\sigma)\mathrm{e}^{-\mu}, & \text{se } y = 0, \\ (1-\sigma)\mathrm{e}^{-\mu}\mu^{y}/y!, & \text{se } y = 1, 2, 3, ..., \end{cases}$$

em que $\mu > 0$ e $0 < \sigma < 1$. A média da distribuição é dada por $(1-\sigma)\mu$, dessa forma μ é um parâmetro de localização da distribuição Poisson inflacionada em zero, e σ é o parâmetro de inflação da probabilidade de contagens nulas.

Como a parametrização original da distribuição apresenta média dada por $(1-\sigma)\mu$, Rigby et al. (2019) apresentam uma reparametrização da distribuição para tornar a interpretação dos coeficientes do modelo sob o GAMLSS mais direta. Nesse caso, o parâmetro μ é reparametrizado como $\mu/(1-\sigma)$. Dessa forma, a função de probabilidade é dada por (RIGBY et al., 2019)

$$f(y \mid \mu, \sigma) = \begin{cases} \sigma + (1-\sigma)\mathrm{e}^{-\left(\frac{\mu}{1-\sigma}\right)}, & \text{se } y = 0, \\ \frac{\mu^{y}}{y!(1-\sigma)^{y-1}}\mathrm{e}^{-\left(\frac{\mu}{1-\sigma}\right)}, & \text{se } y = 1, 2, 3, \ldots, \end{cases}$$

em que $\mu > 0$ e $0 < \sigma < 1$. A média fica então dada pelo próprio parâmetro μ, facilitando interpretações da modelagem. Novamente, o parâmetro σ representa a inflação da probabilidade de contagens nulas.

2.5.4 Geométrica

A distribuição geométrica é uma das distribuições discretas clássicas utilizadas no contexto de regressão para dados de contagem, especialmente quando se deseja modelar sobredispersão leve em relação à distribuição de Poisson. Esta distribuição pertence à família exponencial uniparamétrica, assim pode ser desenvolvida também no contexto dos MLGs. A parametrização da função de probabilidade da distribuição geométrica apresentada por Rigby et al. (2019) é dada por

$$f(y \mid \mu) = \frac{\mu^y}{(\mu+1)^{y+1}}, \quad y \in \{0, 1, 2, ...\}, \quad \mu > 0.$$

A moda da distribuição sempre ocorre em 0 e as probabilidades sempre decaem conforme y aumenta. Essa parametrização oferece a vantagem de fácil interpretabilidade da estimativa do parâmetro μ que representa a própria média da distribuição e variância fica expressada por $\mu + \mu^2$. Nesse sentido, a principal limitação dessa distribuição é sua variância muito rígida definida na forma $\text{Var}(Y) = \mu + \mu^2$, o que restringe sua flexibilidade para acomodar diferentes graus de dispersão, tornando-a inadequada em situações em que a variância observada cresce de forma distinta em relação à média. Cabe mencionar, ainda, que a distribuição admite diferentes parametrizações, aqui foi apresentada a parametrização que expressa a média diretamente em termos do parâmetro μ, cuja sigla no pacote `gamlss` é `GEOM`, o que facilita a interpretação do modelo.

2.5.5 Binomial Negativa

A distribuição Binomial Negativa é amplamente utilizada na modelagem de dados de contagem que apresentem sobredispersão, seja como extensão dos MLGs ou sob o GAMLSS. A função de probabilidade da distribuição é dada por (RIGBY et al., 2019)

$$f(y \mid \mu, \sigma) = \frac{\Gamma(y + \frac{1}{\sigma})}{\Gamma(\frac{1}{\sigma})\Gamma(y+1)} \left(\frac{\sigma\mu}{1+\sigma\mu}\right)^y \left(\frac{1}{1+\sigma\mu}\right)^{1/\sigma}, \quad y \in \{0, 1, 2, ...\},$$

em que $\mu > 0$, $\sigma > 0$ e função gama definida por $\Gamma(t) = \int_0^\infty x^{t-1}e^{-x}\,dx$, com $x > 0$. A distribuição apresenta média $\text{E}(Y) = \mu$ e variância $\text{Var}(Y) = \mu + \sigma\mu^2$, ou seja variância quadrática em μ. Dessa forma, a distribuição Binomial Negativa se apresenta como mais uma alternativa de distribuição para modelar sobredispersão. É importante notar que a distribuição admite diferentes parametrizações, sendo a parametrização `NBI` do pacote `gamlss` a forma usualmente utilizada.

2.5.6 Binomial Negativa Ajustada em Zero

Esse ajuste em zero da distribuição binomial negativa modifica a probabilidade associada ao valor zero, permitindo lidar com situações em que há ausência de zeros ou em que a frequência desse valor não é adequadamente representada pela distribuição de Binomial Negativa tradicional. A função de probabilidade é expressada por (RIGBY et al., 2019)

$$f(y \mid \mu, \sigma, \nu) = \begin{cases} \nu, & \text{se } y = 0, \\ \frac{(1-\nu)}{1-(1+\mu\sigma)^{-1/\sigma}} \cdot \frac{\Gamma\left(y+\nu^{-1}\right) B\left(y+\mu\sigma^{-1}\nu, \sigma^{-1}+\nu^{-1}+1\right)}{\Gamma(y+1)\,\Gamma(\nu^{-1})\,B(\mu\sigma^{-1}\nu, \sigma^{-1}+1)}, & \text{se } y = 1, 2, 3, ..., \end{cases}$$

em que $\mu > 0$, $\sigma > 0$ e $0 < \nu < 1$. O parâmetro μ representa a média da distribuição Binomial Negativa antes do truncamento, σ representa a dispersão da distribuição antes do truncamento, e ν a probabilidade de ocorrência de contagens nulas.

2.5.7 Binomial Negativa Inflacionada em Zero

A distribuição Binomial Negativa inflacionada em zero é uma distribuição discreta adequada para dados de contagem com excesso de zeros. A função de probabilidade é dada por (RIGBY et al., 2019)

$$f(y \mid \mu, \sigma, \nu) = \begin{cases} \nu + (1-\nu)(1+\mu\sigma)^{-1/\sigma}, & \text{se } y = 0, \\ (1-\nu)\dfrac{\Gamma\left(y+\nu^{-1}\right)\, B\left(y+\mu\sigma^{-1}\nu,\, \sigma^{-1}+\nu^{-1}+1\right)}{\Gamma(y+1)\,\Gamma(\nu^{-1})\,B(\mu\sigma^{-1}\nu,\, \sigma^{-1}+1)}, & \text{se } y = 1, 2, 3, ..., \end{cases}$$

em que $\mu > 0$, $\sigma > 0$ e $0 < \nu < 1$. Nesse contexto, a média da distribuição é dada por $(1-\nu)\mu$ e variância é

$$\text{Var}(Y) = \mu(1-\nu) + \mu^2(1-\nu)(\sigma+\nu).$$

Assim, μ é um parâmetro de localização, σ um parâmetro de dispersão, e ν é o parâmetro de inflação da probabilidade de contagens nulas.

2.5.8 Poisson Generalizada

A distribuição de Poisson Generalizada, também conhecida como Poisson Generalizada de Consul, é uma distribuição que apresenta sobredispersão em relação à distribuição Poisson. A função de probabilidade parametrizada em relação à média μ é dada

por (RIGBY et al., 2019)

$$f(y \mid \mu, \sigma) = \left(\frac{\mu}{1+\sigma\mu}\right)^{y} \frac{(1+\sigma y)^{-1}}{y!} \exp\left[\frac{-\mu(1+\sigma y)}{1+\sigma\mu}\right], \quad y \in \{0, 1, 2, ...\},$$

em que $\mu > 0$ e $\sigma > 0$. Nesse caso, μ representa a própria média da distribuição e σ um parâmetro de dispersão. A média e variância dessa reparametrização da distribuição Poisson Generalizada são $\mathrm{E}(Y) = \mu$ e $\mathrm{Var}(Y) = \mu(1+\sigma\mu)^2$. Dessa forma, a distribuição Poisson Generalizada se apresenta como mais uma opção para modelar dados com sobredispersão sob o GAMLSS.

2.5.9 Delaporte

A distribuição Delaporte é utilizada na modelagem de dados com sobredispersão. A função de probabilidade é dada por (RIGBY et al., 2019)

$$f(y \mid \mu, \sigma, \nu) = \frac{\mathrm{e}^{-\mu\nu}}{\Gamma(1/\sigma)} \left[1 + \mu\sigma(1-\nu)\right]^{-1/\sigma} S, \quad y \in \{0, 1, 2, ...\},$$

em que $\mu > 0$, $\sigma > 0$, $0 < \nu < 1$, e S é expressado por

$$S = \sum_{j=0}^{y} \binom{y}{j} \frac{\mu^y \nu^{y-j}}{y!} \left[\mu + \frac{1}{\sigma(1-\nu)}\right]^{-j} \Gamma\left(\frac{1}{\sigma} + j\right).$$

A distribuição tem por média o próprio parâmetro μ e variância igual a $\mu + \mu^2\sigma(1-\nu)^2$, caracterizando a sobredispersão em relação à Poisson. Nesse contexto, o parâmetro μ representa a própria média da distribuição, enquanto os demais parâmetros não possuem interpretação direta.

2.5.10 Família Binomial Negativa

A distribuição da família Binomial Negativa é uma distribuição utilizada para dados de contagem que apresentem sobredispersão. A distribuição possui dois parâmetros relacionados à dispersão, o que confere grande flexibilidade sobre a dispersão para modelagem. Além disso, esta é obtida a partir da reparametrização da distribuição binomial negativa, alterando σ para $\mu^{\nu-2}$. A função de probabilidade é dada por (RIGBY et al., 2019)

$$f(y \mid \mu, \sigma, \nu) = \frac{\Gamma\left(y + \sigma^{-1}\mu^{2-\nu}\right) \sigma^y \mu^{y(\nu-1)}}{\Gamma\left(\sigma^{-1}\mu^{2-\nu}\right) \Gamma(y+1) \left(1+\sigma\mu^{\nu-1}\right)^{\sigma^{-1}\mu^{2-\nu}+y}}, \quad y \in \{0, 1, 2, ...\},$$

em que $\mu > 0$, $\sigma > 0$ e $\nu > 0$. A média da distribuição é dada pelo próprio parâmetro μ e sua variância por $\mu + \sigma\mu^{\nu}$, o que caracteriza a sobredispersão. Assim, μ representa a própria média, σ é parâmetro de dispersão e ν representa o expoente na relação existente entre a variância e a média.

2.5.11 Poisson Dupla

A distribuição Poisson Dupla é uma distribuição extremamente versátil em relação à dispersão, podendo acomodar tanto sobredispersão quanto subdispersão. A função de probabilidade da distribuição Poisson Dupla é dada por (RIGBY et al., 2019)

$$f(y \mid \mu, \sigma) = c(\mu, \sigma)\sigma^{-1/2}e^{-\mu/\sigma}\left(\frac{\mu}{y}\right)^{y/\sigma}\frac{e^{y/\sigma - y}y^{y}}{y!}, \quad y \in \{0, 1, 2, ...\}, \quad \mu > 0, \quad \sigma > 0,$$

em que $c(\mu, \sigma)$ é uma constante normalizadora expressada por

$$c(\mu, \sigma) = \left[\sum_{y=0}^{\infty}\sigma^{-1/2}\mathrm{e}^{-\mu/\sigma}\left(\frac{\mu}{y}\right)^{y/\sigma}\frac{\mathrm{e}^{y/\sigma - y}y^{y}}{y!}\right]^{-1},$$

e adota-se a convenção de que $0^0 = 1$ para o termo y^y quando $y = 0$.

A distribuição Poisson dupla apresenta média aproximadamente igual a μ e variância aproximadamente igual a $\sigma\mu$. Dessa forma, esta apresenta grande flexibilidade em relação à dispersão, sendo adequada tanto para modelagem de dados com sobredispersão, quando $\sigma > 1$, quanto para dados com subdispersão, quando $\sigma < 1$. Além disso, a distribuição Poisson é um caso particular da distribuição Poisson Dupla para $\sigma = 1$.

2.5.12 Burr XII Discreta

A distribuição Burr XII discreta apresenta grande flexibilidade em relação à dispersão, podendo modelar dados tanto com subdispersão quanto sobredispersão. A função de probabilidade é dada por (RIGBY et al., 2019)

$$f(y \mid \mu, \sigma, \nu) = \left[1 + \left(\frac{y}{\mu}\right)^{\sigma}\right]^{-\nu} - \left[1 + \left(\frac{y+1}{\mu}\right)^{\sigma}\right]^{-\nu}, \quad y \in \{0, 1, 2, ...\},$$

em que $\mu > 0$, $\sigma > 0$ e $\nu > 0$. A distribuição tem como média a expressão

$$\mathrm{E}(Y) = \sum_{y=1}^{\infty}\left[1 + \left(\frac{y}{\mu}\right)^{\sigma}\right]^{-\nu},$$

para $\sigma\nu > 1$. Portanto, nenhum dos parâmetros possui interpretação simples, assim, apesar da flexibilidade em relação à dispersão, a distribuição não apresenta simples interpretação prática.

2.5.13 Poisson-Gaussiana Inversa

A distribuição Poisson-Gaussiana Inversa é uma distribuição útil para modelar dados com sobredispersão, além disso, é caso especial da distribuição de Sichel, apresentada mais a frente, quando $\nu = -\frac{1}{2}$. A função de probabilidade é dada por (RIGBY et al., 2019)

$$f(y \mid \mu, \sigma) = \left(\frac{2\alpha}{\pi}\right)^{1/2} \frac{\mu^y \mathrm{e}^{1/\sigma} K_{y-\frac{1}{2}}(\alpha)}{y!\,(\alpha\sigma)^y}, \quad y \in \{0, 1, 2, ...\},$$

em que $\mu > 0$, $\sigma > 0$, $\alpha^2 = \sigma^{-2} + 2\mu\sigma^{-1}$, $\alpha > 0$ e $K_\lambda(t)$ é a função modificada de Bessel de segundo tipo expressada por

$$K_\lambda(t) = \frac{1}{2} \int_0^\infty x^{\lambda-1} \exp\left[-\tfrac{1}{2}t\left(x + x^{-1}\right)\right] dx.$$

A distribuição Poisson-Gaussiana Inversa apresenta média igual a μ e variância igual a $\mu + \sigma\mu^2$, ou seja, o parâmetro μ representa a própria média da distribuição e σ é um parâmetro de dispersão. Assim, como $\mathrm{Var}(Y) > \mathrm{E}(Y)$, a distribuição acomoda sobredispersão. Cabe citar que foi utilizada a parametrização `PIG` do pacote `gamlss` por apresentar relação entre a média e variância mais simples.

2.5.14 Poisson-Gaussiana Inversa Ajustada em Zero

A distribuição Poisson-Gaussiana Inversa ajustada em zero modifica a probabilidade associada ao valor zero, permitindo lidar com situações em que há ausência de zeros ou em que a frequência desse valor não é adequadamente representada pela distribuição de Poisson-Gaussiana Inversa tradicional. A função de probabilidade é expressada por (RIGBY et al., 2019)

$$f(y \mid \mu, \sigma, \nu) = \begin{cases} \nu, & \text{se } y = 0, \\ \dfrac{(1-\nu)}{1 - \exp(1/\sigma - \alpha)} \cdot \left(\dfrac{2\alpha}{\pi}\right)^{1/2} \cdot \dfrac{\mu^y \mathrm{e}^{1/\sigma} K_{y-\frac{1}{2}}(\alpha)}{y!\,(\alpha\sigma)^y}, & \text{se } y = 1, 2, 3, ..., \end{cases}$$

em que $\mu > 0$, $\sigma > 0$, $0 < \nu < 1$ e $\alpha^2 = \sigma^{-2} + 2\mu\sigma^{-1}$ com $\alpha > 0$. O parâmetro μ representa a média da distribuição Poisson-Gaussiana Inversa antes do truncamento, σ a

dispersão da distribuição antes do truncamento, e ν a probabilidade de contagens nulas.

2.5.15 Poisson-Gaussiana Inversa Inflacionada em Zero

A função de probabilidade é dada por (RIGBY et al., 2019)

$$f(y \mid \mu, \sigma, \nu) = \begin{cases} \nu + (1-\nu)\exp(1/\sigma - \alpha), & \text{se } y = 0, \\ (1-\nu) \cdot \left(\dfrac{2\alpha}{\pi}\right)^{1/2} \cdot \dfrac{\mu^y \mathrm{e}^{1/\sigma} K_{y-\frac{1}{2}}(\alpha)}{y!\,(\alpha\sigma)^y}, & \text{se } y = 1, 2, 3, \ldots, \end{cases}$$

em que $\mu > 0$, $\sigma > 0$, $0 < \nu < 1$ e $\alpha^2 = \sigma^{-2} + 2\mu\sigma^{-1}$ com $\alpha > 0$. A média é dada por $(1-\nu)\mu$ e variância é dada por $\mu(1-\nu) + \mu^2(1-\nu)(\sigma+\nu)$. Assim, μ é um parâmetro de localização, σ um parâmetro de dispersão, e ν é o parâmetro de inflação da probabilidade de contagens nulas.

2.5.16 Sichel

A distribuição de Sichel é uma distribuição discreta que apresenta cauda pesada e sobredispersão. A função de probabilidade é dada por (RIGBY et al., 2019)

$$f(y|\mu, \sigma, \nu) = \frac{\mu^y K_{y+\nu}(\alpha)}{y!\,(\alpha\sigma)^{y+\nu}\, K_\nu(1/\sigma)}, \quad y \in \{0, 1, 2, ...\},$$

em que $\mu > 0$, $\sigma > 0$, $-\infty < \nu < \infty$, $\alpha^2 = \sigma^{-2} + 2\mu\sigma^{-1}$, e $K_\lambda(t)$ é a função modificada de Bessel de segundo tipo expressada por

$$K_\lambda(t) = \frac{1}{2}\int_0^\infty x^{\lambda-1} \exp\left[-\tfrac{1}{2}t\left(x + x^{-1}\right)\right] dx.$$

A distribuição possui média dada por $b\mu$ e variância $b\mu + b^2\mu^2 g_1$, em que $b = K_{\nu+1}(1/\sigma)/K_\nu(1/\sigma)$ e $g_1 = 2\sigma(\nu+1)/b + \frac{1}{b^2} - 1$. Nesse contexto, nenhum dos parâmetros apresenta interpretação direta.

Rigby et al. (2019) apresentam uma reparametrização da distribuição de Sichel, de tal forma que μ como μ/b. Assim, a média da distribuição seria o próprio parâmetro μ. A função de probabilidade fica dada por (RIGBY et al., 2019)

$$f(y \mid \mu, \sigma, \nu) = \frac{(\mu/b)^y K_{y+\nu}(\alpha)}{y!\,(\alpha\sigma)^{y+\nu}\, K_\nu(1/\sigma)}, \quad y \in \{0, 1, 2, ...\},$$

em que $\mu > 0$, $\sigma > 0$, $-\infty < \nu < \infty$, $\alpha^2 = \sigma^{-2} + 2\mu(b\sigma)^{-1}$, $b = K_{\nu+1}(1/\sigma)/K_\nu(1/\sigma)$ e

$K_\lambda(t)$ é a função modificada de Bessel de segundo tipo expressada por

$$K_\lambda(t) = \frac{1}{2}\int_0^\infty x^{\lambda-1}\exp\left[-\tfrac{1}{2}t\left(x+x^{-1}\right)\right]\,dx.$$

A média dessa reparametrização é dada pelo próprio parâmetro μ e variância por $\mu + \mu^2 g_1$. Portanto, percebe-se que nessa parametrização a interpretação de μ é direta, porém, os demais parâmetros permanecem sem interpretação.

2.5.17 Sichel Ajustada em Zero

A função de probabilidade é dada por (RIGBY et al., 2019)

$$f(y \mid \mu,\sigma,\nu,\tau) = \begin{cases} \tau, & \text{se } y = 0, \\ \dfrac{1-\tau}{1-\frac{K_\nu(\alpha)}{(\alpha\sigma)^\nu K_\nu(1/\sigma)}} \cdot \dfrac{\mu^y K_{y+\nu}(\alpha)}{y!\,(\alpha\sigma)^{y+\nu}\,K_\nu(1/\sigma)}, & \text{se } y = 1,2,3,\ldots, \end{cases}$$

em que $\mu > 0$, $\sigma > 0$, $-\infty < \nu < \infty$, $0 < \tau < 1$, $\alpha^2 = \sigma^{-2} + 2\mu(b\sigma)^{-1}$, $b = K_{\nu+1}(1/\sigma)/K_\nu(1/\sigma)$ e $K_\lambda(t)$ é a função modificada de Bessel de segundo tipo. A média da distribuição é dada por $c\mu$. Portanto, μ representa um parâmetro de localização da distribuição, τ representa a probabilidade de ocorrência de contagens nulas e os demais parâmetros não apresentam interpretação direta.

2.5.18 Sichel Inflacionada em Zero

A função de probabilidade é dada por (RIGBY et al., 2019)

$$f(y \mid \mu,\sigma,\nu,\tau) = \begin{cases} \tau + (1-\tau)\dfrac{K_\nu(\alpha)}{(\alpha\sigma)^\nu K_\nu(1/\sigma)}, & \text{se } y = 0, \\ (1-\tau)\dfrac{\mu^y K_{y+\nu}(\alpha)}{y!\,(\alpha\sigma)^{y+\nu}\,K_\nu(1/\sigma)}, & \text{se } y = 1,2,3,\ldots, \end{cases}$$

em que $\mu > 0$, $\sigma > 0$, $-\infty < \nu < \infty$ e $0 < \tau < 1$. A média da distribuição é dada por $(1-\tau)\mu$. Assim, o parâmetro μ é um parâmetro de localização da distribuição, τ é o parâmetro de inflação da probabilidade de ocorrência de contagens nulas, e os demais parâmetros não possuem interpretação direta.

2.5.19 Beta-Binomial Negativa

A distribuição Beta-Binomial Negativa, também chamada de Beta-Pascal ou Waring Generalizada, apresenta cauda pesada e sobredispersão. Vale citar, ainda, que a

distribuição de Waring é caso especial da distribuição Beta-Binomial Negativa, quando $\nu = 1$. A função de probabilidade é dada por (RIGBY et al., 2019)

$$f(y \mid \mu, \sigma, \nu) = \frac{\Gamma\left(y + \nu^{-1}\right)\, B\left(y + \mu\sigma^{-1}\nu,\, \sigma^{-1} + \nu^{-1} + 1\right)}{\Gamma(y+1)\,\Gamma(\nu^{-1})\, B(\mu\sigma^{-1}\nu,\, \sigma^{-1} + 1)}, \quad y \in \{0, 1, 2, ...\},$$

em que $\mu > 0$, $\sigma > 0$ e $\nu > 0$. A média da distribuição é dada pelo próprio parâmetro μ e sua variância por

$$\mathrm{Var}(Y) = \begin{cases} \mu(1+\mu\nu)(1+\sigma\nu^{-1})(1-\sigma)^{-1}, & \text{se } \sigma < 1, \\ \infty, & \text{se } \sigma \geq 1. \end{cases}$$

Assim, tem-se sobredispersão em relação à distribuição de Poisson e cauda pesada, especialmente para valores elevados de σ. Dessa forma, o parâmetro μ representa a própria média, enquanto os demais parâmetros não possuem interpretação direta.

2.5.20 Beta-Binomial Negativa Ajustada em Zero

A distribuição Beta-Binomial Negativa ajustada em zero é uma distribuição discreta adequada para dados de contagem. Esse ajuste modifica a probabilidade associada ao valor zero, permitindo lidar com situações em que há ausência de zeros ou em que a frequência desse valor não é adequadamente pela distribuição original. A função de probabilidade é expressada por (RIGBY et al., 2019)

$$f(y \mid \mu, \sigma, \nu, \tau) = \begin{cases} \tau, & \text{se } y = 0, \\ \frac{1-\tau}{1-\frac{B(\mu\sigma^{-1}\nu, \sigma^{-1}+\nu^{-1}+1)}{B(\mu\sigma^{-1}\nu, \sigma^{-1}+1)}} \frac{\Gamma\left(y+\nu^{-1}\right) B\left(y+\mu\sigma^{-1}\nu, \sigma^{-1}+\nu^{-1}+1\right)}{\Gamma(y+1)\,\Gamma(\nu^{-1})\, B(\mu\sigma^{-1}\nu, \sigma^{-1}+1)}, & \text{se } y = 1, 2, 3, \ldots, \end{cases}$$

em que $\mu > 0$, $\sigma > 0$, $\nu > 0$ e $0 < \tau < 1$. Assim, μ representa um parâmetro de localização da distribuição e o parâmetro τ representa a probabilidade de ocorrência de contagens nulas, enquanto os demais parâmetros não apresentam interpretação direta.

2.5.21 Beta-Binomial Negativa Inflacionada em Zero

A função de probabilidade é dada por (RIGBY et al., 2019)

$$f(y \mid \mu, \sigma, \nu, \tau) = \begin{cases} \tau + (1-\tau)\dfrac{B(\mu\sigma^{-1}\nu, \sigma^{-1}+\nu^{-1}+1)}{B(\mu\sigma^{-1}\nu, \sigma^{-1}+1)}, & \text{se } y = 0, \\ (1-\tau)\dfrac{\Gamma\left(y + \nu^{-1}\right)\, B\left(y + \mu\sigma^{-1}\nu,\, \sigma^{-1} + \nu^{-1} + 1\right)}{\Gamma(y+1)\,\Gamma(\nu^{-1})\, B(\mu\sigma^{-1}\nu,\, \sigma^{-1} + 1)}, & \text{se } y = 1, 2, 3, \ldots, \end{cases}$$

em que $\mu > 0$, $\sigma > 0$ e $\nu > 0$. Nesse sentido, a média da distribuição é expressada por $(1 - \tau)\mu$ e variância por

$$(1 - \tau)\mu(1 + \mu\nu)(1 + \sigma\nu^{-1})(1 - \sigma)^{-1} + \tau(1 - \tau)\mu^2.$$

Dessa forma, μ é interpretado como um parâmetro de localização, e τ como o parâmetro de inflação de inflação da probabilidade de contagens nulas, porém, os demais parâmetros não apresentam interpretação direta.

2.5.22 Yule

Uma distribuição adequada para a modelagem de dados de contagem é a distribuição Yule, que é ideal para modelagem de dados com cauda pesada. Cabe citar, também, que a distribuição é um caso especial da distribuição Waring, que é citada mais a frente, quando $\mu = \sigma$. A função de probabilidade é dada por (RIGBY et al., 2019)

$$f(y \mid \mu) = (\mu^{-1} + 1)B(y + 1, \mu^{-1} + 2), \quad y \in \{0, 1, 2, ...\}, \quad \mu > 0,$$

em que $B(x, y) = \int_0^1 t^{x-1}(1 - t)^{y-1}dt$ é a função Beta. A distribuição apresenta moda sempre em zero e as probabilidades decaem conforme y aumenta. A distribuição apresenta $\mathrm{E}(Y) = \mu$, e a variância é dada por

$$\mathrm{Var}(Y) = \begin{cases} \mu(\mu + 1)^2(1 - \mu)^{-1}, & \text{se } \mu < 1, \\ \infty, & \text{se } \mu \geq 1. \end{cases}$$

Assim, a distribuição Yule apresenta cauda pesada especialmente para valores altos de μ.

2.5.23 Waring

A distribuição Waring, também conhecida como Beta-Geométrica, é uma distribuição cuja principal particularidade é sua cauda pesada, que a torna útil para modelar dados que apresentem número considerável de contagens altas. A função de probabilidade é dada por (RIGBY et al., 2019)

$$f(y \mid \mu, \sigma) = \frac{B(y + \mu\sigma^{-1},\, \sigma^{-1} + 2)}{B(\mu\sigma^{-1},\, \sigma^{-1} + 1)}, \quad y \in \{0, 1, 2, ...\},$$

em que $\mu > 0$, $\sigma > 0$ e $B(x, y) = \int_0^1 t^{x-1}(1-t)^{y-1}dt$ é a função Beta. A distribuição apresenta média igual ao próprio parâmetro μ e variância dada por

$$\text{Var}(Y) = \begin{cases} \mu(\mu+1)(1+\sigma)(1-\sigma)^{-1}, & \text{se } \sigma < 1, \\ \infty, & \text{se } \sigma \geq 1, \end{cases}$$

em que o parâmetro σ não é interpretável. A distribuição tem cauda pesada principalmente para valores elevados de σ.

2.5.24 Logarítmica

A distribuição logarítmica é uma das distribuições discretas adequadas para dados de contagem que apresentem cauda pesada e ausência de zeros. Cabe citar que a distribuição, por pertencer à família exponencial uniparamétrica, também pode ser desenvolvida no contexto dos MLGs. A função de probabilidade é dada por (RIGBY et al., 2019)

$$f(y \mid \mu) = \frac{-[\log(1-\mu)]^{-1}\mu^y}{y}, \quad y \in \{1, 2, 3, ...\}, \quad 0 < \mu < 1.$$

A distribuição logarítmica se mostra adequada para modelagem de dados com sobredispersão, pois $\text{Var}(Y) = \alpha\mu(1-\alpha\mu)(1-\mu)^{-2}$ e $\text{E}(Y) = \alpha\mu(1-\mu)^{-1}$, porém ainda apresenta certa rigidez na variância limitando sua flexibilidade. Além disso, em razão do parâmetro μ não ter interpretações.

O parâmetro μ não tem interpretação clara, ou seja, não representa nem média, mediana ou qualquer medida de tendência central, o que limita ainda mais a utilidade prática dessa distribuição para dados de contagem.

2.5.25 Logarítmica Ajustada em Zero

O ajuste em zero da distribuição logarítmica amplia o domínio da distribuição original, incorporando explicitamente a ocorrência do valor zero, o que possibilita modelar situações em que há presença de zeros nos dados. Além disso, mantém-se a cauda pesada da distribuição. A função de probabilidade é dada por (RIGBY et al., 2019)

$$f(y \mid \mu, \sigma) = \begin{cases} \sigma, & \text{se } y = 0, \\ (1-\sigma)(\alpha\mu^y)/y, & \text{se } y = 1, 2, 3, \ldots, \end{cases}$$

em que $0 < \mu < 1$, $0 < \sigma < 1$ e $\alpha = -[\log(1-\mu)]^{-1}$. A média da distribuição fica dada por $(1-\sigma)\alpha\mu(1-\mu)^{-1}$, ou seja, novamente o parâmetro μ não apresenta interpretação direta. Já o parâmetro σ representa a probabilidade de contagens nulas. Assim, em razão da interpretabilidade bem limitada do parâmetro μ sob o GAMLSS, essa distribuição não apresenta grande utilidade prática.

2.5.26 Zipf

A distribuição Zipf, também conhecida como distribuição zeta de Riemann ou distribuição discreta de Pareto, é uma distribuição discreta, com cauda pesada, adequada para dados de contagem. A função de probabilidade é dada por (RIGBY et al., 2019)

$$f(y \mid \mu) = \frac{y^{-(\mu+1)}}{\zeta(\mu+1)}, \quad y \in \{1,2,3,...\}, \quad \mu > 0,$$

em que $\zeta(b) = \sum_{i=1}^{\infty} i^{-b}$ é a função zeta de Riemann. A moda sempre ocorre em 1 e as probabilidades decaem conforme y aumenta. A distribuição acomoda bem dados com valores mais extremos e assimetria por ser uma distribuição que apresenta cauda pesada, especialmente quando μ se aproxima de zero. No entanto, a estimativa do parâmetro μ é de difícil interpretação, visto que a média da distribuição é dada por

$$\mathrm{E}(Y) = \begin{cases} \frac{\zeta(\mu)}{\zeta(\mu+1)} & \text{se } \mu > 1 \\ \infty & \text{se } \mu \leq 1 \end{cases}$$

Assim, apesar de sua flexibilidade para modelar dados com cauda pesada, a utilidade prática da distribuição Zipf é limitada pela dificuldade de interpretação dos seus coeficientes no contexto da modelagem estatística sob o GAMLSS.

2.5.27 Zipf Ajustada em Zero

O ajuste em zero da distribuição Zipf amplia o domínio da distribuição original, incorporando explicitamente a ocorrência do valor zero. Além disso, mantém-se a cauda pesada da distribuição. A função de probabilidade é dada por (RIGBY et al., 2019)

$$f(y \mid \mu, \sigma) = \begin{cases} \sigma, & \text{se } y = 0, \\ (1-\sigma)y^{-(\mu+1)}/\zeta(\mu+1), & \text{se } y = 1,2,3,..., \end{cases}$$

em que $\mu > 0$, $0 < \sigma < 1$ e $\zeta(b) = \sum_{i=1}^{\infty} i^{-b}$ é a função zeta de Riemann. A média da distribuição fica dada por

$$\mathrm{E}(Y) = \begin{cases} b(1-\sigma), & \text{se } \mu > 1, \\ \infty, & \text{se } \mu \leq 1, \end{cases}$$

em que $b = \zeta(\mu)/\zeta(\mu+1)$. Assim, o parâmetro μ não tem interpretação direta, já o parâmetro σ representa a probabilidade de contagens nulas.

2.5.28 Síntese das distribuições de contagem

Dada a variedade de distribuições disponíveis para modelar dados de contagem, foi elaborado um diagrama das distribuições apresentadas separadas pelas características. Assim, a Figura 1 resume as características principais das distribuições mencionadas anteriormente. No diagrama foram utilizadas as siglas das distribuições presentes no pacote `gamlss` apresentadas na Tabela 2. Além disso, as versões ajustadas em zero das distribuições Zipf e logarítmica incorporam o zero ao seu domínio, diferentemente das versões originais, que não incluem o zero.

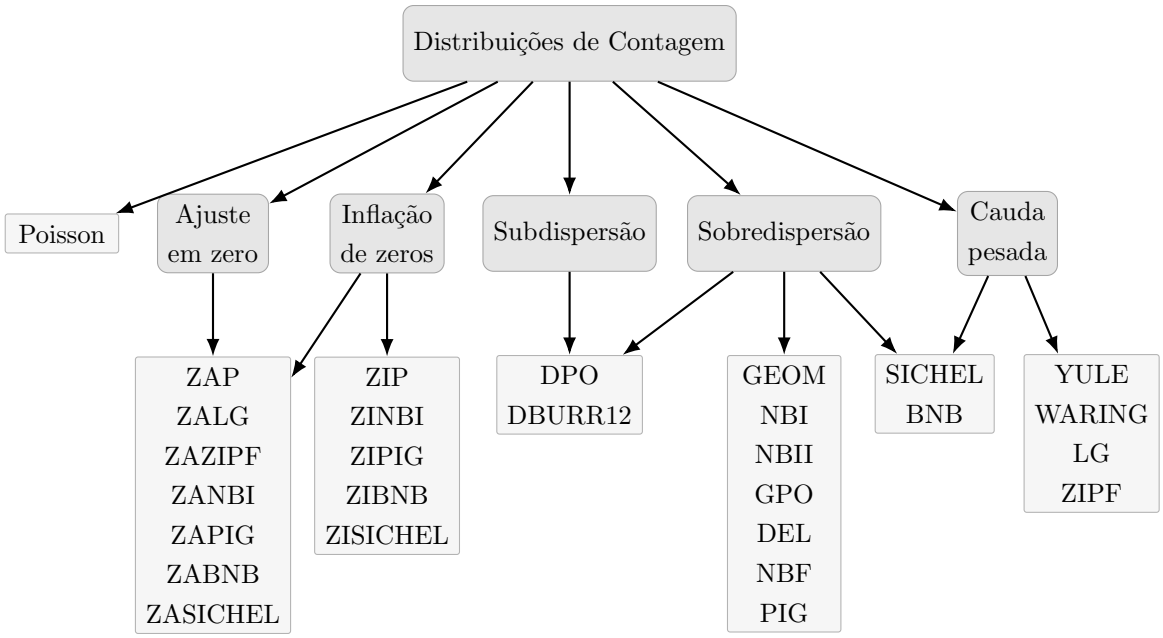


Figura 1: Diagrama de classificação das distribuições de contagem.

2.6 Métodos de análise diagnóstica e avaliação da qualidade do ajuste

Nesta seção, serão abordados alguns métodos utilizados na análise diagnóstica de MLGs e GAMLSSs. A análise diagnóstica é uma etapa importante para garantir que os resultados obtidos a partir dos modelos sejam confiáveis e verificar a adequação do ajuste do modelo aos dados.

2.6.1 Medida de alavancagem

Nos Modelos Lineares Generalizados, os pontos de alavanca são observações que possuem grande influência sobre os valores ajustados pelo modelo, devido à sua posição no espaço das covariáveis. Assim como no modelo linear clássico, o conceito de alavancagem está relacionado à matriz de projeção dos valores ajustados (PAULA, 2025). Contudo, nos MLGs, essa matriz depende da função de ligação e da variância da resposta, o que torna o cálculo da alavancagem mais complexo e dependente da especificação do modelo.

A alavancagem é geralmente medida pela diagonal da matriz de alavancagem $\hat{H}$, representada por $\hat{h}_{ii}$, que leva em conta os pesos das observações (matriz $\hat{W}$) e a estrutura do preditor linear. A forma geral da matriz de alavancagem $\hat{H}$ nos MLGs é dada por

$$\hat{H} = \hat{W}^{\frac{1}{2}} X \left(X^\top \hat{W} X \right)^{-1} X^\top \hat{W}^{\frac{1}{2}},$$

em que $\hat{W} = \text{diag}\{\hat{w}_1, \ldots, \hat{w}_n\}$ é a matriz diagonal das estimativas dos pesos que depende da variância da distribuição da resposta, e $\hat{w}_i$ é a estimativa do peso correspondente à i-ésima observação. As medidas de alavancagem $\hat{h}_{ii}$, que dizem respeito à diagonal principal de $\hat{H}$, variam entre 0 e 1 e possuem média igual a $(k+1)/n$, onde $k+1$ é o número de parâmetros do modelo e n o número de observações. Sob um modelo bem ajustado, espera-se que essas medidas apresentem valores baixos. A avaliação da alavancagem é usualmente realizada por meio do gráfico de $\hat{h}_{ii}$ pelos índices das observações, onde podem ser destacados os pontos que excedem valores como $2(k+1)/n$ (duas vezes a média amostral) ou $3(k+1)/n$ (três vezes a média amostral).

2.6.2 Resíduos

Entre as principais ferramentas utilizadas na análise diagnóstica de modelos de regressão estão os resíduos. Em MLGs, os resíduos podem ser definidos de diversas formas,

cada uma com propósitos e propriedades distintas. Paula (2025) apresenta os resíduos de Pearson studentizados, os resíduos desvio e os resíduos quantílicos (DUNN; SMYTH, 1996). Os resíduos de Pearson nos MLGs são análogos aos resíduos padronizados do modelo linear normal. Eles são definidos por

$$r_{p_i} = \frac{\sqrt{\hat{\phi}}\,(y_i - \hat{\mu}_i)}{\sqrt{\widehat{V}_i}}, \qquad i = 1, 2, \ldots, n,$$

em que $\widehat{V}_i = V(\hat{\mu}_i)$ é a estimativa da função de variância e $\hat{\phi}$ é a estimativa do parâmetro de dispersão. Para obter resíduos com variância aproximadamente igual a 1, introduz-se uma correção que considera a medida de alavancagem $\hat{h}_{ii}$. Assim, o resíduo de Pearson studentizado é dado por

$$t_{S_i} = \frac{\sqrt{\hat{\phi}}\,(y_i - \hat{\mu}_i)}{\sqrt{\widehat{V}_i(1 - \hat{h}_{ii})}}.$$

Já em relação ao resíduo desvio, um dos resíduos mais utilizados em MLGs (PAULA, 2025), é definido a partir dos componentes da função desvio expressada na forma

$$t_{D_i} = \frac{d^*(y_i; \hat{\mu}_i)}{\sqrt{1 - \hat{h}_{ii}}} = \frac{\sqrt{\hat{\phi}} d(y_i; \hat{\mu}_i)}{\sqrt{1 - \hat{h}_{ii}}},$$

em que $d(y_i; \hat{\mu}_i) = \pm\sqrt{2\left\{y_i(\hat{\theta}_i - \tilde{\theta}_i) + \left[b(\tilde{\theta}_i) - b(\hat{\theta}_i)\right]\right\}}$, em que θ é o parâmetro canônico da família exponencial, $\tilde{\theta}_i$ é o valor de θ_i no modelo saturado, $\hat{\theta}_i$ é o valor de θ_i no modelo ajustado, $b(\theta)$ é a função de ligação canônica e o sinal de $d(y_i; \hat{\mu}_i)$ é o mesmo de $(y_i - \hat{\mu}_i)$. São especialmente relevantes para avaliar a qualidade do ajuste em MLGs, pois valores elevados indicam observações que o modelo ajusta mal. Em geral, a distribuição dos resíduos desvio não é conhecida, mas espera-se que assumam valores próximos de zero sob um bom ajuste.

Os resíduos quantílicos, particularmente úteis sob os GAMLSS, são construídos com base na inversa da função de distribuição acumulada do modelo ajustado e têm uma adaptação para o caso discreto como é mostrado por Dunn e Smyth (1996). Se y é uma observação de uma variável resposta discreta, como é o caso de dados de contagem, então a função de distribuição acumulada $F(y \mid \boldsymbol{\theta})$ apresenta saltos, pois y é discreto. Nesse caso, os resíduos quantílicos normalizados aleatorizados ficam definidos por

$$\hat{t}_{Q_i} = \Phi^{-1}(\tilde{u}_i),$$

em que $\Phi^{-1}(\cdot)$ é a função quantil da distribuição normal padrão e $\tilde{u}_i$ é uma variável aleatória gerada de uma distribuição uniforme no intervalo

$$\tilde{u}_i \sim \mathbb{U}\left(\hat{F}(y_i - 1 \mid \hat{\boldsymbol{\theta}}), \hat{F}(y_i \mid \hat{\boldsymbol{\theta}})\right).$$

Aqui, $\hat{F}(y_i \mid \hat{\boldsymbol{\theta}})$ representa a função de distribuição acumulada ajustada avaliada em y_i e $\hat{F}(y_i - 1 \mid \hat{\boldsymbol{\theta}})$ é o valor da função de distribuição acumulada imediatamente anterior. Sob a correta especificação do modelo, os resíduos $\hat{t}_{Q_i}$ têm distribuição aproximadamente normal padrão e são independentes, para n grande. Além disso, dado que o modelo está bem ajustado, o resíduo quantílico é o único que a distribuição assintótica é conhecida.

2.6.3 Técnicas gráficas

A partir dos resíduos existem diversas formas de avaliação das suposições dos modelos que podem ser realizadas graficamente. Diante disso, o uso dos resíduos apresentados podem variar, porém, suas interpretações para avaliação dos pressupostos, no geral, são semelhantes. Dessa forma, a seção apresentará alguns exemplos de técnicas gráficas para análise diagnóstica de MLGs e GAMLSSs. Assim, um dos métodos para avaliação da qualidade do ajuste é o gráfico de probabilidade normal com envelope simulado, também chamado gráfico quantil-quantil ou *QQ-plot* com envelope simulado, que permite avaliar se a distribuição suposta para a resposta está adequada. No gráfico, os resíduos ordenados são comparados aos quantis esperados de uma distribuição normal padrão. Como os resíduos não são exatamente normais em amostras finitas, o envelope simulado fornece um referencial para a variação esperada. Esse envelope é construído a partir de várias réplicas da amostra geradas sob o modelo ajustado. Quando os resíduos observados permanecem em sua maioria dentro das bandas de confiança, a um dado nível de significância, e seguem um padrão aproximadamente linear, há indícios de que o modelo está coerente com os dados. Além disso, no contexto de dados de contagem, o modo como os pontos cruzam a reta de referência oferece indicação sobre a adequação da variância: quando o cruzamento ocorre de forma crescente, indica existência de sobredispersão em relação à Poisson na variável resposta, enquanto um cruzamento decrescente sugere subdispersão (PAULA, 2025).

A Figura 2 apresenta os resíduos desvio, também chamados resíduos componente do desvio, com envelopes simulados de um modelo de regressão Poisson ajustado a dados referentes a um certo conjunto de dados de contagem de exemplo. Nos MLGs, a análise diagnóstica baseia-se, principalmente, na análise dos resíduos, e nesses casos, normalmente

opta-se pelos resíduos desvio ou pelos resíduos quantílicos.

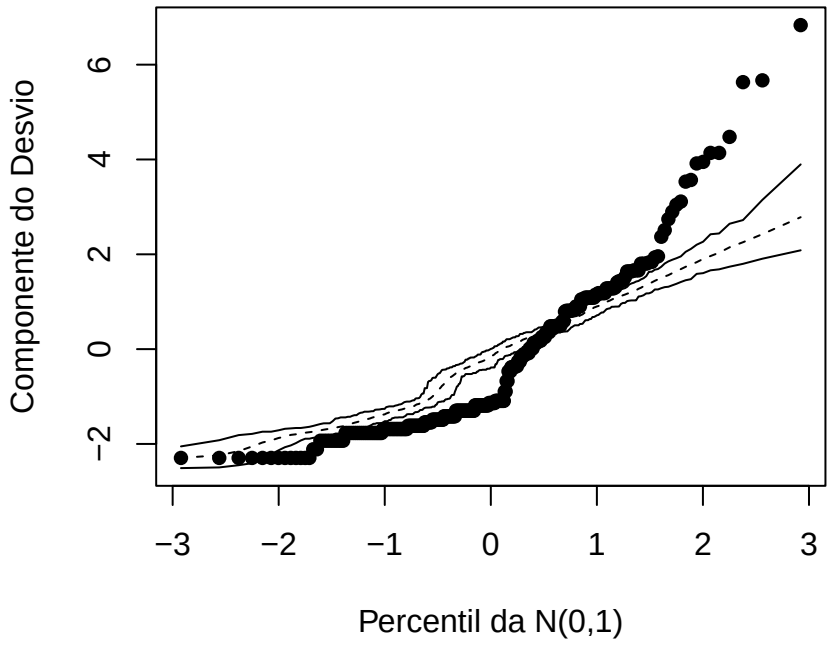


Figura 2: Gráfico de probabilidade normal com envelope simulado para os resíduos desvio do modelo de regressão Poisson ajustado.

Diante disso, a partir da Figura 2, pode-se observar que os pontos cortam o envelope de forma crescente, o que indica que um modelo com distribuição com sobredispersão seria mais adequado. Assim, podem-se citar o modelo de regressão binomial negativo, o GAMLSS com distribuição Poisson dupla, entre outros cujas distribuições se adéquem a dados com sobredispersão em relação à Poisson.

Para modelos mais flexíveis como os GAMLSS, que permitem a modelagem de múltiplos parâmetros da distribuição, a análise diagnóstica também costuma considerar resíduos mais generalizados, como os resíduos quantílicos. Essas ferramentas são muitas vezes acompanhadas por gráficos de diagnóstico específicos, como os gráficos de *worm plot* (BUUREN; FREDRIKS, 2001), que avaliam a adequação da distribuição da resposta. O *worm plot* pode ser entendido como um *QQ-plot* sem tendência, em que os valores dos resíduos quantílicos são subtraídos da reta de referência. No eixo horizontal apresentam-se os valores esperados dos resíduos quantílicos dado o modelo bem ajustado, enquanto no eixo vertical são exibidos os resíduos quantílicos centrados nos respectivos quantis teóricos da distribuição normal padrão. O gráfico mostra pontos que refletem o grau de

afastamento entre os resíduos observados e seus valores teóricos aproximados, além de incluir bandas de confiança de 95% para cada ponto e a curva ajustada representada por um polinômio cúbico ajustado aos valores do *worm plot*, o que facilita a identificação de padrões sistemáticos. Sob o modelo bem ajustado, espera-se que 95% dos pontos se encontrem dentro das bandas de confiança e que a curva ajustada não indique uma clara tendência entre os pontos.

A Tabela 3 evidencia as interpretações sobre tendências nos *worm plots* e suas reflexões na média, variância, assimetria e curtose das distribuições ajustadas.

Tabela 3: 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 Superestimado
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 Figura 3 apresenta o gráfico de *worm plot* do modelo de regressão normal ajustado sob o GAMLSS, em que os pontos representam os resíduos quantílicos das observações pelos quantis teóricos, as linhas pontilhadas delimitam as bandas de confiança de 95% e a linha vermelha representa a curva ajustada associada ao polinômio cúbico que auxilia na identificação da tendência.

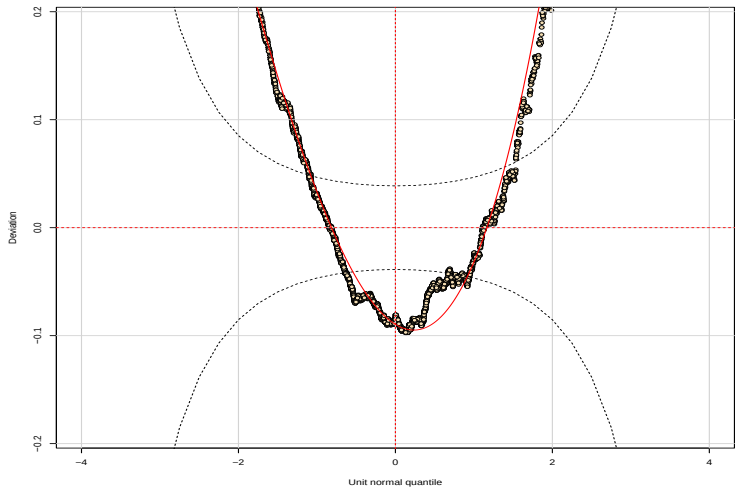


Figura 3: *Worm plot* do modelo de exemplo

Pelo *worm plot* do modelo, pode-se notar que os pontos cortam as bandas de confiança, o que indica um ajuste ruim do modelo, além disso, o formato de U exibido pelos pontos mostra que a distribuição normal usada no modelo apresenta excesso de assimetria à esquerda, de acordo com Tabela 3, e deve-se procurar uma distribuição mais assimétrica à direita que a normal para um ajuste mais adequado.

Além do *worm plot*, outros gráficos, que utilizam os resíduos quantílicos, são elaborados para avaliar a qualidade do ajuste dos modelos. A Figura 4 apresenta os gráficos de análise de diagnóstico dos resíduos quantílicos para verificação da adequabilidade de outro modelo de exemplo, comumente utilizados sob o GAMLSS.

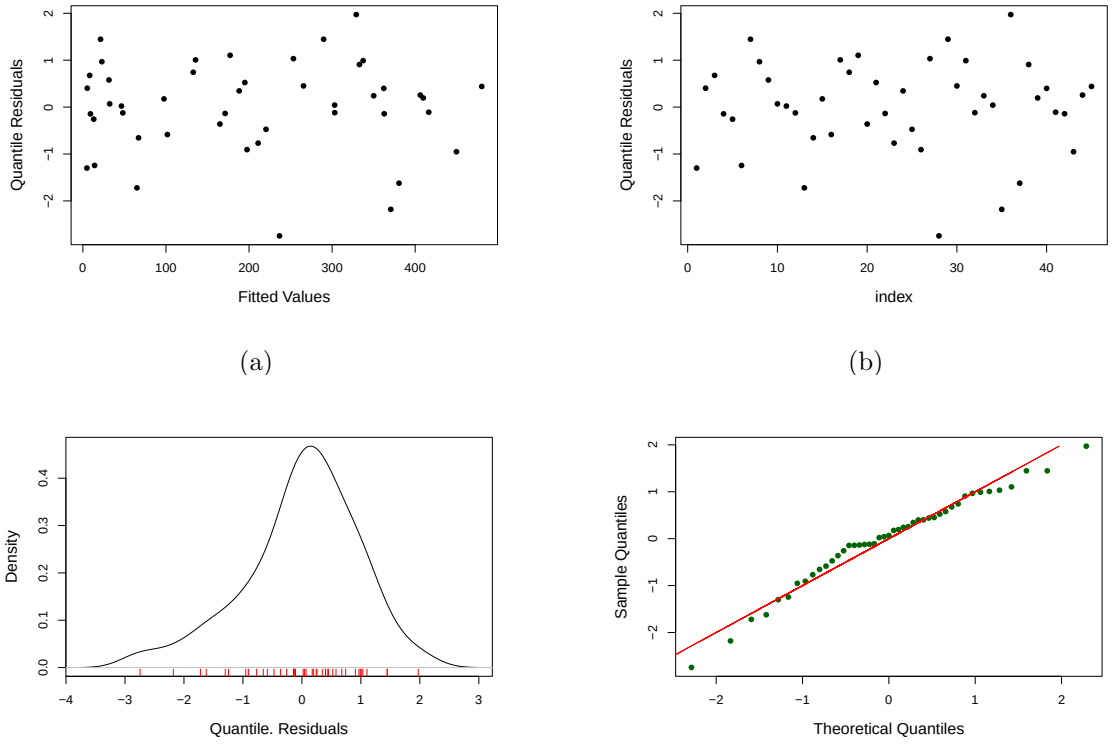


Figura 4: Gráficos de diagnóstico para os resíduos quantílicos sob o modelo de exemplo: (a) dispersão dos resíduos pelos valores ajustados de μ; (b) dispersão dos resíduos pelos índices das observações; (c) densidade estimada dos resíduos; (d) *QQ-plot* dos resíduos.

A partir da figura 4 (a), espera-se do modelo bem ajustado que a dispersão dos pontos se mantenha homogênea ao longo dos valores ajustado, o que se observa no caso do modelo de exemplo. Além disso, pela Figura 4 (b), espera-se que os pontos não apresentem qualquer padrão no gráfico o que indica que não há presença clara de correlação entre os resíduos e os índices, nesse sentido, os pontos não apresentam padrão, ou seja, não há indícios contra o bom ajuste do modelo. Em contrapartida, a Figura

4 (c) mostra uma densidade estimada assimétrica à esquerda dos resíduos, o que é um indicativo de inadequação do modelo, pois, espera-se do modelo com bom ajuste que os pontos apresentem comportamento simétrico e semelhante, até certo ponto, à densidade da distribuição normal padrão, em razão do uso dos resíduos quantílicos. Por fim, a partir da Figura 4 (d), nota-se que diversos pontos se afastam moderadamente da reta de referência, o que indica violação da normalidade dos resíduos quantílicos. Pelo *QQ-plot*, espera-se que os pontos apresentem boa aderência à reta de referência com pouca distanciação entre os pontos e a reta, desvios da reta indicam inadequação da distribuição ajustada.

3 Aplicações

Neste capítulo são apresentadas aplicações a dados reais dos modelos de regressão para dados de contagem apresentados no Capítulo 3. Dessa forma, este capítulo tem como objetivo ilustrar o uso dos GAMLSS. Os conjuntos de dados selecionados abrangem diversos padrões de variabilidade presentes em dados de contagem, permitindo ilustrar como cada modelo lida com variação da dispersão, excesso de zeros, entre outras características. Em cada aplicação são apresentados o ajuste, a análise diagnóstica e a interpretação dos parâmetros dos modelos. Os *scripts* computacionais utilizados para a implementação das aplicações e obtenção dos resultados apresentados neste capítulo estão disponíveis em um repositório público no *GitHub*, acessível em: https://github.com/MPMassaro/TCC-Regressao-de-Contagem.

Para as aplicações computacionais dos modelos apresentados, foram utilizados os seguintes pacotes do `R`: `gamlss` (STASINOPOULOS et al., 2024), para o ajuste dos GAMLSSs.

No desenvolvimento computacional do trabalho, diversas funções do `R` foram empregadas para ajuste, diagnóstico e visualização dos modelos considerados. Entre as principais, destacam-se:

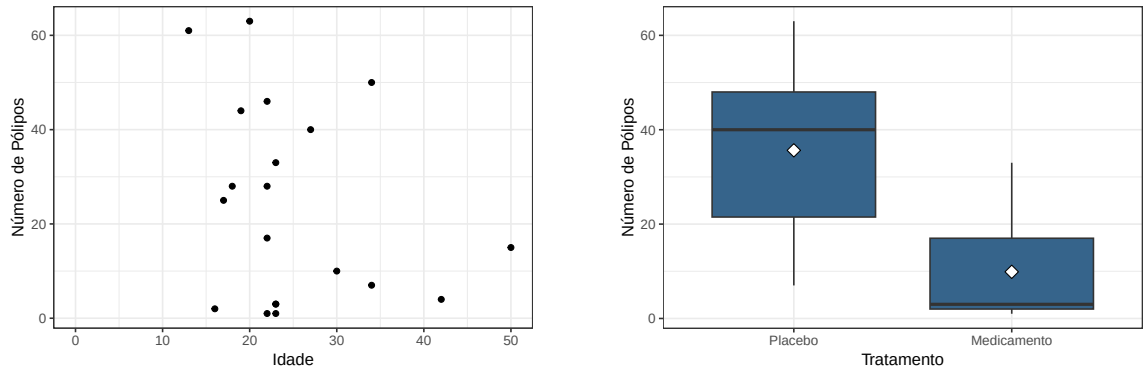
- `gamlss`: utilizada para o ajuste dos modelos GAMLSSs, permitindo especificar distribuições, preditores e funções de ligação para cada parâmetro.
- `fitDist`: utilizada para ajustar marginalmente múltiplas distribuições candidatas e comparar seus desempenhos segundo critérios de ajuste.
- `wp`: empregada para gerar os *worm plots* dos resíduos quantílicos, utilizados na avaliação da qualidade do ajuste.
- `plot`: utilizada para gerar os gráficos de diagnóstico sob o modelo utilizando os resíduos quantílicos.

3.1 Aplicação 1: Número de pólipos em função da idade e do tipo de tratamento

Para apresentar a primeira aplicação dos modelos de regressão de contagem estudados foi utilizado o banco `polyps` disponível no pacote `GLMsData` (DUNN; SMYTH, 2022) no software `R`. O banco é baseado em um estudo realizado em 1993 acerca do uso

de sulindaco no tratamento de polipose adenomatosa familiar, uma doença hereditária caracterizada pela formação de pólipos no intestino grosso que podem evoluir para um câncer (GIARDIELLO et al., 1993). Nesse sentido, o banco apresenta informações acerca do número de pólipos de 20 pacientes, além da idade e se o tratamento foi conduzido usando um placebo ou o medicamento. A modelagem por regressão do número de pólipos permite avaliar o efeito da idade e verificar se o uso do medicamento altera significativamente essa contagem, além de mensurar a intensidade da diminuição da quantidade de lesões.

A Figura 5 mostra o gráfico de dispersão do número de pólipos do paciente pela idade e o *boxplot* da variável resposta por tipo de tratamento. No *boxplot*, as médias de cada grupo estão representadas pelos quadrados brancos sobrepostos. A partir da Figura 5 (a), nota-se que pouco menos da metade dos pacientes apresentou menos de 20 lesões, e que as contagens mais elevadas tendem a ocorrer em indivíduos mais jovens, no geral, com menos de 30 anos de idade. Na Figura 5 (b), pode-se ver que o número de pólipos entre os paciente que receberam tratamento com sulindaco é muito inferior aos daqueles que receberam o placebo apresentando média abaixo de 10 e mediana igual a 3. Em contrapartida, os pacientes que receberam placebo apresentaram média em torno de 36 lesões e mediana de 40.



(a) (b)

Figura 5: (a) Gráfico de dispersão entre o número de pólipos e idade em anos do paciente; (b) *boxplot* do número de pólipos e tipo de tratamento utilizado.

A Tabela 4 apresenta os valores do Critério de Informação de Akaike (AIC) para o ajuste marginal das distribuições discretas de contagem ao número de lesões dos pacientes. A Tabela 4 oferece uma indicação inicial de quais distribuições podem ser adequadas para modelar a variável resposta, o que é bastante útil devido ao elevado número de distribuição adequadas para dados de contagem disponíveis no pacote `gamlss`. É importante salientar, no entanto, que, ao utilizar o AIC marginal produzido pela função `fitDist` do pacote,

Tabela 4: Valores do critério de informação AIC para diferentes distribuições discretas de contagem ajustadas marginalmente para o número de pólipos do paciente.

Distribuição	AIC	Distribuição	AIC	Distribuição	AIC	Distribuição	AIC
GEOM	170,0257	SI	173,7871	ZINBF	176,0240	YULE	223,0566
GEOMo	170,0257	SICHEL	173,7871	ZIBNB	176,0240	PO	468,9958
ZANBI	171,6661	NBF	174,0240	PIG	176,4387	ZAP	470,9958
NBII	172,0240	ZINBI	174,0240	ZALG	176,5128	ZIP2	470,9958
NBI	172,0240	BNB	174,0245	ZAPIG	177,6999	ZIP	470,9958
WARING	172,0257	GPO	174,4331	ZIPIG	178,4387		
DPO	172,2617	LG	174,5128	ZIPF	189,0087		
DEL	173,1569	ZISICHEL	175,7871	ZAZIPF	191,0087		

deve-se ter cautela, pois os modelos comparados não são necessariamente encaixados e o critério reflete apenas o ajuste marginal. Como a distribuição condicional às covariáveis pode ser substancialmente distinta da distribuição marginal variável resposta, decisões baseadas apenas no valor do AIC marginal podem levar a escolhas inadequadas.

Inicialmente, será ajustado o modelo mais simples usualmente aplicado a dados de contagem, no caso o GAMLSS com distribuição Poisson, denominado Modelo 1, com função de ligação logarítmica para modelar a media, desde que o parâmetro seja positivo, que é usualmente utilizada para parâmetros cujo espaço paramétrico é positivo por ter a interpretação mais simples. Como sugerido pela Tabela 4, não se espera um bom desempenho desse modelo, já que este apresentou um dos maiores valores de AIC marginal. Para avaliar a qualidade do ajuste do modelo de regressão Poisson considerando as covariáveis, denominado Modelo 1, foram calculados os resíduos quantílicos e construídos os gráficos de diagnóstico para o modelo, apresentados na Figura 6.

Pela Figura 6 (a), pode-se perceber uma leve diminuição da variância dos resíduos quantílicos conforme aumenta-se o valor ajustado para μ, o que pode indicar um desvio das suposições do modelo, dado que se espera variância constante sob o modelo bem ajustado. Porém, cabe ressaltar que o tamanho reduzido da amostra pode contribuir para esse comportamento heterocedástico dos resíduos quantílicos. A partir da Figura 6 (b), não há indicação de qualquer padrão, ou seja, não há presença clara de correlação dos resíduos aos índices, porém, é possível observar um ponto aberrante com valor do resíduo muito elevado, o que pode prejudicar o ajuste do modelo. Pela Figura 6 (c), que apresenta a densidade estimada dos resíduos e espera-se que se assemelhe à densidade da distribuição normal, percebe-se um comportamento bem divergente do esperado da distribuição normal, o que é reforçado pela Figura 6 (d) que mostra clara falta de aderência dos pontos à reta de referência, o que indica forte desvio da normalidade dos resíduos quantílicos. Cabe citar, ainda, que o comportamento dos pontos na Figura 6 (d), cruzando a linha de referência de forma crescente, indica sobredispersão da variável resposta.

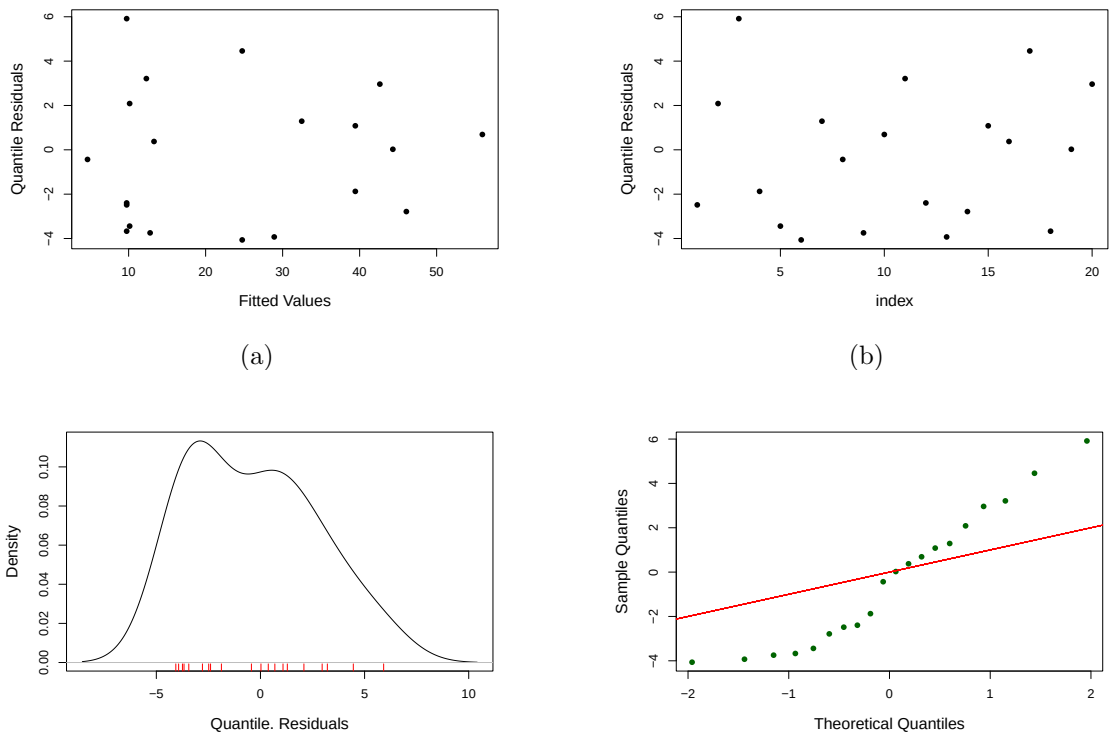


Figura 6: Gráficos de diagnóstico para os resíduos quantílicos sob o Modelo 1 da Aplicação 1: (a) dispersão dos resíduos pelos valores ajustados de μ; (b) dispersão dos resíduos pelos índices das observações; (c) densidade estimada dos resíduos; (d) *QQ-plot* dos resíduos.

A Figura 7 apresenta o *worm plot* dos resíduos quantílicos do Modelo 1 para avaliar o ajuste do modelo.

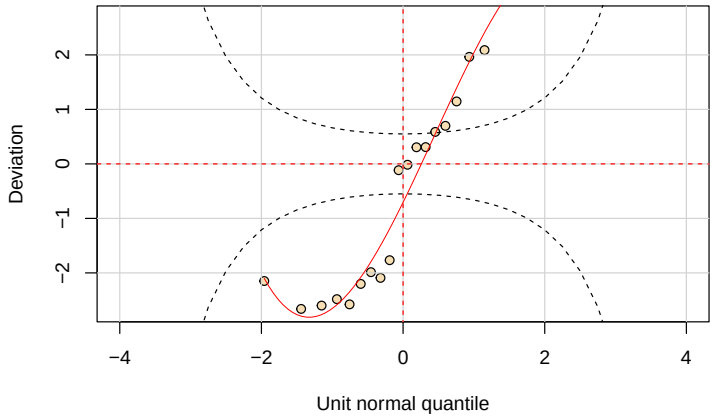


Figura 7: *Worm plot* dos resíduos quantílicos sob o Modelo 1 da Aplicação 1.

Pelo Figura 7, percebe-se uma tendência crescente dos pontos cruzando as bandas

de confiança, além de a maioria dos pontos se encontrar fora das bandas de confiança, indicando que o modelo de regressão Poisson não está bem ajustado aos dados. Além disso, o comportamento da reta no *worm plot* de inclinação positiva reforça o que pôde-se interpretar do *QQ-plot*, o que indica a necessidade de uma distribuição com sobredispersão em relação à Poisson de acordo com as informações apresentadas na Tabela 3 da Seção 2.6.3. Vale ressaltar que esse comportamento persistiu quando os resíduos quantílicos foram recalculados, lembrando que os resíduos quantílicos para modelos discretos são aleatorizados.

Em seguida, foi ajustado um GAMLSS utilizando a distribuição geométrica, que apresenta sobredispersão em relação à Poisson e pode possivelmente melhor ajustar os dados. O GAMLSS sob distribuição geométrica, denotado por Modelo 2, será ajustado utilizando novamente a função de ligação logarítmica, usual para os parâmetros cujo domínio é estritamente positivo. A fim de avaliar a qualidade do ajuste do modelo, foram calculados os resíduos quantílicos e construídos os gráficos de diagnóstico, mostrados na Figura 8.

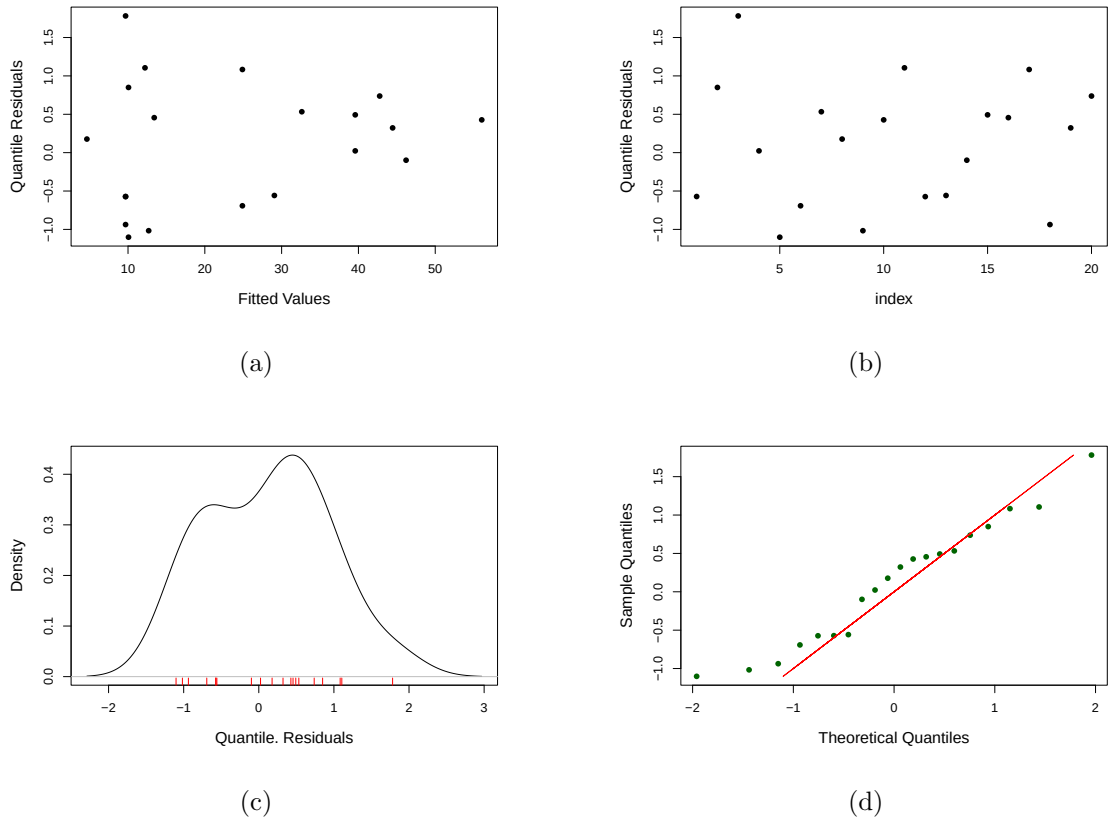


Figura 8: Gráficos de diagnóstico para os resíduos quantílicos sob o Modelo 2 da Aplicação 1: (a) dispersão dos resíduos pelos valores ajustados de μ; (b) dispersão dos resíduos pelos índices das observações; (c) densidade estimada dos resíduos; (d) *QQ-plot* dos resíduos.

Pela Figura 8 (a), nota-se novamente uma aparente diminuição da variância dos resíduos quantílicos conforme os valores ajustados se elevam, o que indica um possível desvio da estrutura de variância $\mu + \mu^2$ da geométrica, porém, há de se considerar o tamanho da amostra pequeno que possivelmente contribui para isso. Em contrapartida, pela Figura 8 (b), não se percebe qualquer tendência entre os pontos, o que indica que não há desvio da suposição de independência das respostas. As Figuras 8 (c) e 8 (d) mostram uma leve melhora em relação ao desvio da normalidade dos resíduos em comparação ao Modelo 1, porém, alguns desvios ainda são perceptíveis. A partir do *QQ-plot*, nota-se alguns pontos que se distanciam um pouco da reta de referência.

Para aprofundar a análise diagnóstica, foi elaborado o *worm plot* do Modelo 2, apresentado na Figura 9. Pelo gráfico, todos os pontos se encontram dentro das bandas de contagem, no entanto, é perceptível uma leve tendência decrescente dos pontos indicando um excesso de variância do modelo. Cabe citar, ainda, que a tendência permaneceu mesmo quando os resíduos quantílicos foram recalculados. Assim, considerando as inadequabilidades do modelo e as informações obtidas a partir do *worm plot*, deve-se optar por um modelo ajustado sob uma distribuição com menos variância que a geométrica, mas que ainda seja sobredisperso em relação à Poisson.

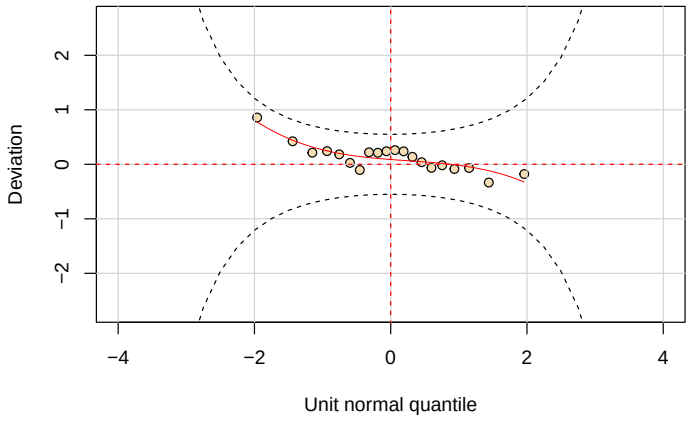


Figura 9: *Worm plot* dos resíduos quantílicos sob o Modelo 2 da Aplicação 1.

Foi-se, então, ajustado um GAMLSS sob distribuição binomial negativa, que apresenta maior flexibilidade acerca da variância por meio da modelagem do parâmetro de dispersão, que será denominado Modelo 3. Após testar diferentes combinações de covariáveis e funções de ligação para os parâmetros da distribuição, decidiu-se por modelar μ, que representa a média, em função de ambas as covariáveis e utilizando função de ligação logarítmica usual. Por sua vez, σ, um parâmetro de dispersão da distribuição, foi

modelado apenas em relação ao tipo de tratamento, também com função de ligação logarítmica, pois essa especificação apresentou ajuste razoavelmente adequado. Para avaliar a qualidade do ajuste desse modelo, foram elaborados gráficos de diagnóstico baseados no resíduo quantílico, apresentados na Figura 10. A partir da Figura 10 (a), observa-se uma melhora considerável em relação ao comportamento heterocedástico evidente para o mesmo gráfico do Modelo 2. No entanto, é possível perceber um aumento da variância entre os valores ajustados 20 e 30, seguido por uma redução da variância em valores mais elevados dos ajustes. Pela Figura 10 (b), não há qualquer tipo de padrão entre os pontos do gráfico, ou seja, nenhum indício de desvio da suposição de não correlação das respostas, além de não haver ponto com resíduo quantílico muito distante de zero. Na Figura 10 (c), nota-se que a densidade estimada dos resíduos quantílicos aparenta ser simétrica, próxima da simetria esperada de uma distribuição normal. Por fim, pela Figura 10 (d), não é observado nenhum ponto que se distancie de forma expressiva da linha de tendência, assim, não há indicativo de desvio do pressuposto de normalidade dos resíduos quantílicos no Modelo 3.

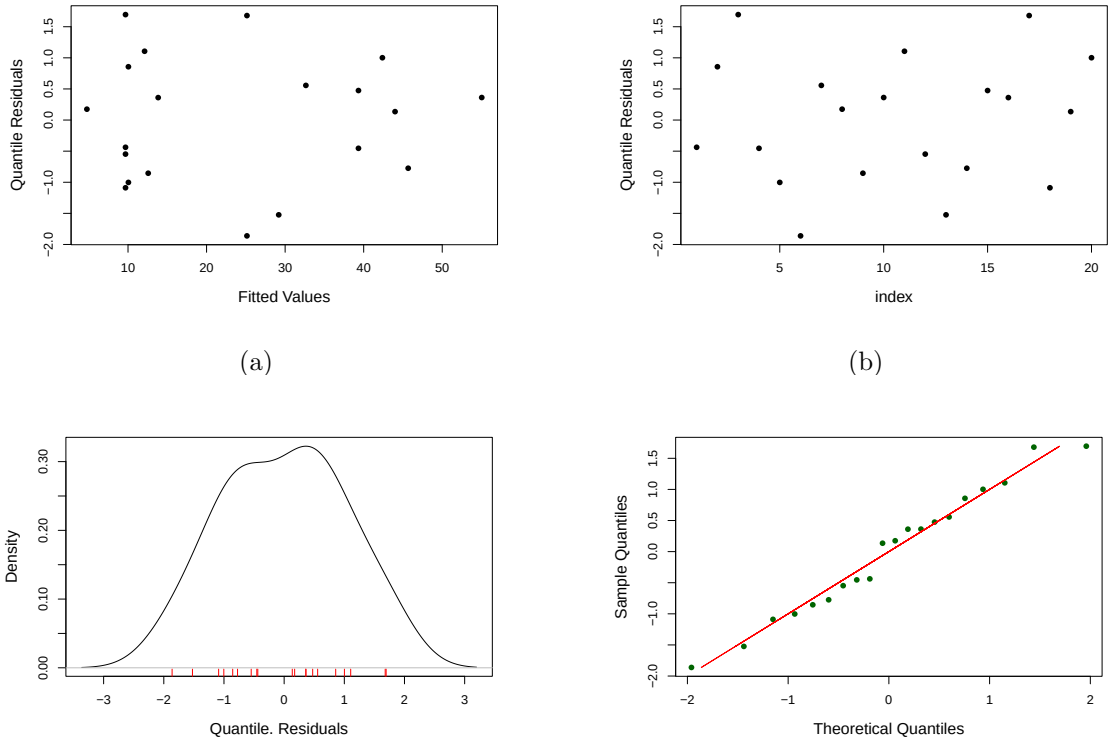


Figura 10: Gráficos de diagnóstico para os resíduos quantílicos sob o Modelo 3 da Aplicação 1: (a) dispersão dos resíduos pelos valores ajustados de μ; (b) dispersão dos resíduos pelos índices das observações; (c) densidade estimada dos resíduos; (d) *QQ-plot* dos resíduos.

A Figura 11 apresenta o *worm plot* do Modelo 3 a fim de avaliar a adequação do ajuste do modelo. Nesse sentido, a partir do gráfico pode-se observar que todos os pontos se encontraram dentro das bandas de confiança e próximos de zero, sem nenhuma tendência aparente, o que indica boa adequação do modelo. Assim, como o Modelo 3 foi eficaz em acomodar a sobredispersão dos dados e apresentou, no geral, boa adequação com relação aos resíduos quantílicos, com ressalva para um comportamento moderado de heterocedasticidade observado na Figura 10 (a), optou-se por selecionar esse modelo para apresentar as estimativas dos coeficientes e as respectivas interpretações.

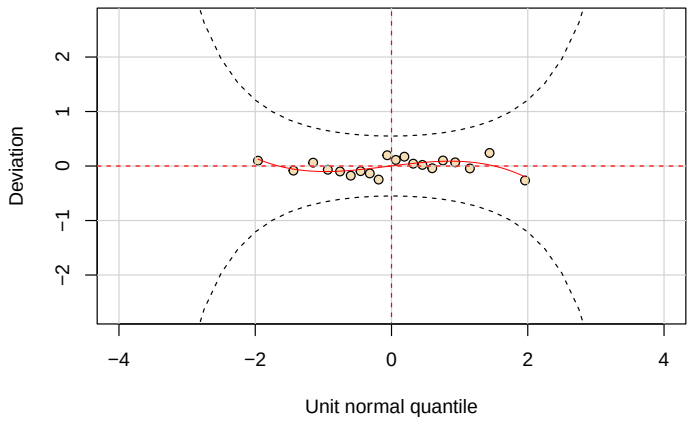


Figura 11: *Worm plot* dos resíduos quantílicos sob o Modelo 3 da Aplicação 1.

A Tabela 5 apresenta as estimativas dos coeficientes, os erros-padrão, as estatísticas z e p-valores dos testes de Wald sob o Modelo 3. Cabe citar, ainda, que $\beta_{\mu 0}$, $\beta_{\mu 1}$ e $\beta_{\mu 2}$ representam, respectivamente, o intercepto em relação à μ, o coeficiente associado à idade e o coeficiente associado ao uso de medicamento, sendo o uso de placebo como referencial. Os parâmetros $\beta_{\sigma 0}$ e $\beta_{\sigma 1}$ são, respectivamente, o intercepto associado ao submodelo do parâmetro σ e o coeficiente associado ao tratamento com uso de medicamento. Cabe citar, ainda, que σ é um parâmetro de dispersão e μ é a própria média da distribuição binomial negativa.

Tabela 5: Estimativas, erros-padrão, estatísticas z e p-valores dos coeficientes sob o Modelo 3 da Aplicação 1.

Coeficiente	Estimativa	Erro-Padrão	Estatística z	p-valor
$\beta_{\mu 0}$	4,4928	0,4154	10,815	<0,001
$\beta_{\mu 1}$	-0,0373	0,0150	-2,487	0,013
$\beta_{\mu 2}$	-1,3647	0,4056	-3,365	<0,001
$\beta_{\sigma 0}$	-1,5101	0,4936	-3,059	0,002
$\beta_{\sigma 1}$	1,6505	0,6820	2,420	0,015

A partir dos valores das estatísticas de teste e p-valores apresentados na Tabela

5, considerando um nível de significância de $\alpha = 0{,}10$, todos os coeficientes associados aos submodelos de μ e σ foram significativamente diferentes de zero. Assim, ambas as covariáveis se mostraram relevantes para explicar o número médio de lesões nos pacientes com a doença e o tipo de tratamento foi relevante para explicar a dispersão das lesões.

Considerando que o ajuste do Modelo 3 é adequado, as estimativas são confiáveis e interpretáveis. Em razão do uso da função de ligação logarítmica no submodelo de μ, a interpretação das estimativas é dada pela função inversa de seus valores, ou seja, a exponencial dos valores das estimativas dos coeficientes. Nesse sentido, no submodelo de μ, a exponencial do intercepto corresponde ao valor da média quando todas as covariáveis são fixadas em zero. As exponenciais dos coeficientes das covariáveis representam fatores multiplicativos sobre esse valor, indicando a mudança relativa esperada na média para um aumento unitário na covariável (ou para a comparação com a categoria de referência, no caso de variáveis qualitativas). Assim, para o intercepto do submodelo de μ, tem-se que $\mathrm{e}^{\hat{\beta}_{\mu 0}} = \mathrm{e}^{4{,}4928} \approx 89{,}37$, porém, esse valor não é interpretável, pois não é coerente fixar a variável idade em zero. Em relação à $\hat{\beta}_{\mu 1}$, o valor negativo da estimativa indica que o aumento da idade está associado à redução do número de lesões. Para interpretar esse coeficiente, calcula-se $\mathrm{e}^{\hat{\beta}_{\mu 1}} = \mathrm{e}^{-0{,}0373} \approx 0,963$, ou seja, o número de pólipos diminui em média 3,7% com a adição de um ano à idade do paciente. Além disso, assim como ocorre na idade, o valor negativo de $\hat{\beta}_{\mu 2}$ indica que o uso do medicamento diminui o número de lesões se comparado ao placebo. A interpretação é obtida a partir de $\mathrm{e}^{\hat{\beta}_{\mu 2}} = \mathrm{e}^{-1{,}3647} \approx 0,255$, o que representa uma diminuição de 74,5% do número médio de lesões em relação aos pacientes que recebem o placebo. Para exemplificar, um paciente com polipose, com 30 anos e que está recebendo o medicamento apresentaria em média 7,46[1] pólipos enquanto um paciente com a mesma idade que recebe o placebo apresenta em média 29,19[2] pólipos.

Por fim, as interpretações das estimativas dos coeficientes do submodelo de σ são dadas de forma semelhante ao submodelo de μ, em razão do uso da mesma função de ligação logarítmica. A exponencial do intercepto fornece o valor de σ quando as covariáveis são fixadas em zero, enquanto as exponenciais dos coeficientes associados às covariáveis representam fatores multiplicativos sobre esse valor de referência. A exponencial da estimativa do intercepto $\hat{\beta}_{\sigma 0}$, $\mathrm{e}^{-1{,}5101} \approx 0{,}221$, representa o valor de σ para os pacientes que receberam o placebo. Já o coeficiente relativo ao tratamento, $\hat{\beta}_{\sigma 1} = 1{,}6505$, implica um efeito multiplicativo de $\mathrm{e}^{1{,}6505} \approx 5{,}21$ sobre o parâmetro de dispersão, ou seja, o uso do medicamento aumenta σ em aproximadamente 421% em comparação ao placebo, sugerindo maior variabilidade no número de lesões entre os pacientes tratados.

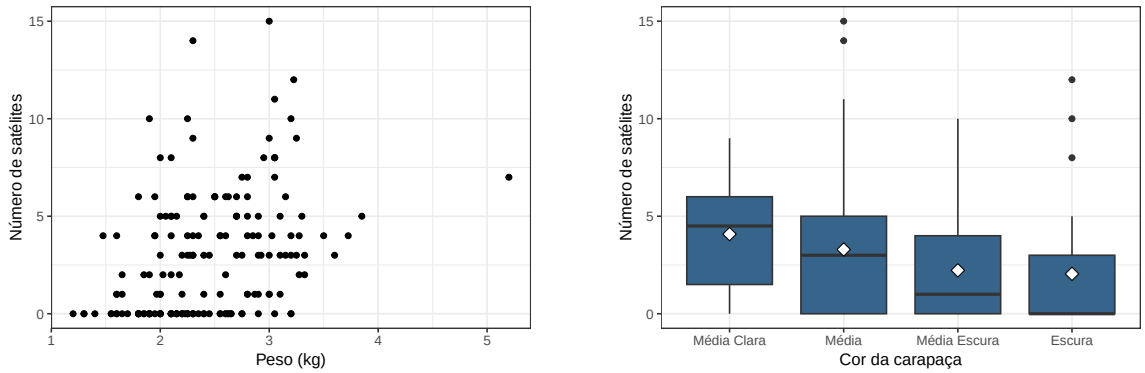
[1] Valor obtido a partir do cálculo $\mathrm{e}^{4{,}4928} \cdot \mathrm{e}^{30 \cdot (-0{,}0373)} \cdot \mathrm{e}^{-1{,}3647} \approx 7{,}46$

[2] Valor obtido a partir do cálculo $\mathrm{e}^{4{,}4928} \cdot \mathrm{e}^{30 \cdot (-0{,}0373)} \approx 29{,}19$

3.2 Aplicação 2: Número de satélites em função das características das fêmeas de caranguejo-ferradura

A segunda aplicação refere-se ao número de satélites de fêmeas de caranguejo-ferradura. O banco `hcrabs` disponível no pacote `GLMsData` (DUNN; SMYTH, 2022) do `R` é proveniente de um estudo observacional sobre reprodução de caranguejos-ferradura através da observação do número de machos que acompanham uma fêmea da espécie, denominados satélites (BROCKMANN, 1996). Nesse sentido, o banco contém informações de 173 fêmeas de caranguejo-ferradura sobre o número de satélites que acompanham a fêmea, a largura da carapaça da fêmea em centímetros, o peso da fêmea em quilogramas, a coloração da carapaça e condição das espinhas do caranguejo. Assim, a modelagem por regressão do número de satélites de fêmeas de caranguejo ferradura pode ajudar a identificar possíveis fatores que aumentam ou diminuem essa contagem. Nesse sentido, devido à elevada correlação entre o peso e a largura da carapaça, optou-se por não incluir a variável largura nos modelos, uma vez que o peso apresentou maior capacidade explicativa e, portanto, foi selecionado para permanecer nas análises. Além disso, a variável referente à condição das espinhas foi desconsiderada por apresentar pouca variação entre os grupos, contribuindo de forma limitada para a modelagem.

A Figura 12 apresenta o gráfico de dispersão do número de satélites pelo peso, e o *boxplot* da variável resposta pela coloração da carapaça. No *boxplot*, as médias de cada grupo estão representadas pelos quadrados brancos sobrepostos.



(a) (b)

Figura 12: Gráfico de dispersão entre o número de satélites e (a) peso do caranguejo em quilogramas e *boxplot* do número de satélites pela (b) cor da carapaça.

Pela Figura 12 (a), observa-se uma suave tendência de aumento no número de satélites conforme o peso da fêmea cresce. Entretanto, nota-se também uma quantidade

expressiva de contagens nulas, indicando a presença de um número considerável de caranguejos sem satélites. Nesse sentido, é possível que um modelo sob distribuição Poisson performe mal em razão do excesso de zeros, além disso, o excesso de contagens nulas tende a causar sobredispersão em relação à Poisson. A partir da Figura 12 (b), nota-se uma diminuição gradual do número de satélites quanto mais escura a carapaça da fêmea. Além disso, apenas as colorações média e escura apresentam algumas observações extremas.

A Tabela 6 apresenta os valores do AIC segundo as distribuições discretas de contagem ajustadas marginalmente ao número de satélites das fêmeas de caranguejo-ferradura.

Tabela 6: Valores do critério de informação AIC para diferentes distribuições discretas de contagem ajustadas marginalmente para o número de satélites de caranguejo-ferradura.

Distribuição	AIC	Distribuição	AIC	Distribuição	AIC	Distribuição	AIC
LG	581,6562	ZASICHEL	746,7032	GEOM	772,1208	GPO	782,0245
ZIPF	648,5624	ZIBNB	746,7032	GEOMo	772,1208	PIG	797,2416
ZINBI	744,7032	DPO	756,9750	SICHEL	773,4092	ZALG	809,4147
ZANBI	744,7032	ZIP	767,2292	NBF	773,4092	YULE	838,2788
ZIPIG	744,9700	ZAP	767,2292	SI	773,4092	ZAZIPF	876,3209
ZAPIG	744,9700	ZIP2	767,2292	DEL	773,4092	PO	990,0893
ZINBF	746,7032	NBI	771,4092	BNB	773,4093		
ZISICHEL	746,7032	NBII	771,4092	WARING	774,1207		

A partir da Tabela 6, nota-se que distribuições com sobredispersão em relação à Poisson e inflacionadas/ajustadas em zero apresentaram menores valores de AIC marginal. Isso sugere que distribuições com essas características podem melhor acomodar os dados, porém, deve-se considerar o AIC marginal com cautela. Nesse sentido, como os modelos comparados não são necessariamente encaixados, pode-se desconsiderar as duas distribuições com menor AIC marginal, logarítmica e Zipf, que não possuem o zero no domínio e, portanto, não são adequadas para modelagem dos dados em questão.

Inicialmente foi ajustado um GAMLSS sob distribuição Poisson, denominado Modelo 1, e utilizada a função de ligação logarítmica, a função de ligação usual para o parâmetro μ que é estritamente positivo. Dessa forma, a Figura 13 apresenta os gráficos de diagnóstico utilizando os resíduos quantílicos com o objetivo de verificar a qualidade do ajuste do modelo. A partir da Figura 13 (a), percebe-se uma grande concentração dos pontos no canto esquerdo do gráfico sem indícios claros de heterocedasticidade entre os resíduos quantílicos. Pela Figura 13 (b), não é possível observar qualquer tendência entre os pontos, ou seja, nenhum indicativo contra a pressuposto de independência das respostas, porém, existem pontos com valores de resíduo muito elevados e que podem comprometer o bom ajuste do modelo. Na Figura 13 (c), nota-se grave desvio da normalidade dos resíduos quantílicos que não se distribuem de forma simétrica ao redor de zero

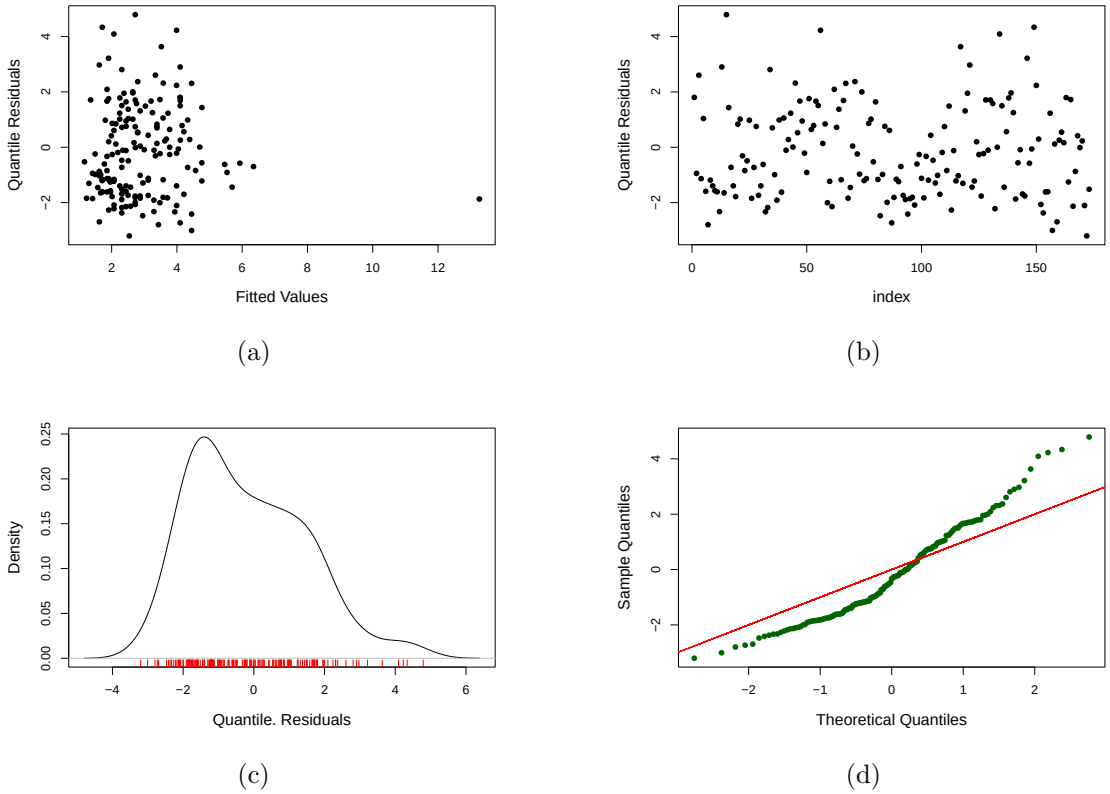


Figura 13: Gráficos de diagnóstico para os resíduos quantílicos sob o Modelo 1 da Aplicação 2: (a) dispersão dos resíduos pelos valores ajustados de μ; (b) dispersão dos resíduos pelos índices das observações; (c) densidade estimada dos resíduos; (d) *QQ-plot* dos resíduos.

e cuja densidade se distancia bastante da simetria esperada. Por fim, a partir da Figura 13 (d), percebe-se que os pontos não acompanham a linha de tendência e a cruzam de forma crescente, indicando claramente uma sobredispersão em relação à distribuição de Poisson. A fim de avaliar a qualidade do ajuste, foi elaborado também um *worm plot* dos resíduos quantílicos sob o modelo, apresentado na Figura 14.

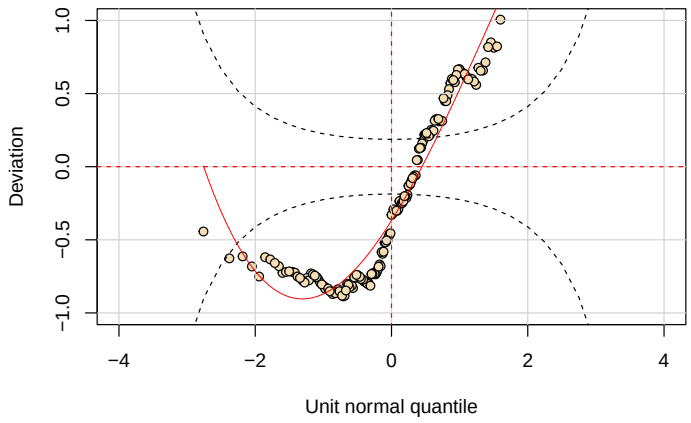


Figura 14: *Worm plot* dos resíduos quantílicos sob o Modelo 1 da Aplicação 2.

Pela Figura 14, vê-se que a maioria dos pontos se encontra fora das bandas de confiança, indicando forte inadequabilidade do modelo. Além disso, pode-se ver uma forte tendência de inclinação positiva entre os pontos, o que sugere que um modelo com sobredispersão pode acomodar melhor os dados. No entanto, deve-se levar em consideração também o excesso de zeros na variável resposta que pode contribuir para esse comportamento de sobredispersão observado. Cabe citar, ainda, que esse comportamento se repetiu ao se recalcular os resíduos quantílicos do modelo.

Dessa forma, foi ajustado um modelo com distribuição Poisson inflacionada em zero, denominado Modelo 2. Foram avaliadas diversas combinações de covariáveis e funções de ligação nos submodelos dos parâmetros, e decidiu-se por modelar μ em função de ambas as covariáveis consideradas, utilizando a função de ligação logarítmica. Para o submodelo do parâmetro de inflação de zeros σ, escolheu-se um modelo apenas em função da variável peso, empregando a função de ligação logito, que é a escolha usual para parâmetros definidos no intervalo unitário. Foram construídos então os gráficos de diagnóstico do Modelo 2 utilizando os resíduos quantílicos, apresentados na Figura 15.

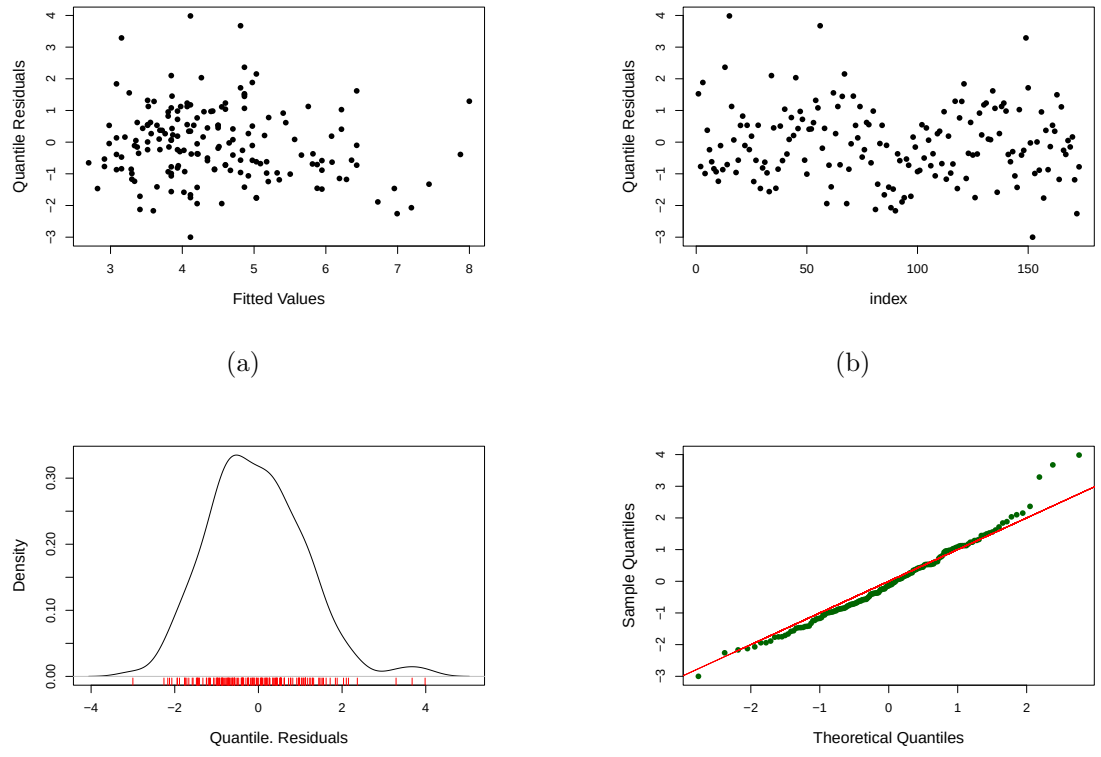


Figura 15: Gráficos de diagnóstico para os resíduos quantílicos sob o Modelo 2 da Aplicação 2: (a) dispersão dos resíduos pelos valores ajustados de μ; (b) dispersão dos resíduos pelos índices das observações; (c) densidade estimada dos resíduos; (d) *QQ-plot* dos resíduos.

Pela Figura 15 (a), ao desconsiderar os 3 pontos destoantes que apresentam valor do resíduo bastante elevado, percebe-se, no geral, um comportamento homoscedástico dos resíduos quantílicos em relação aos valores ajustados. A partir da Figura 15 (b), não se nota padrão entre os resíduos e os índices, assim, não há indicativo de desvio do pressuposto de não correlação entre as respostas. Pelas Figuras 15 (c) e 15 (d), percebe-se grande melhora da normalidade dos resíduos quantílicos quando comparada ao Modelo 1. Porém, pelo *QQ-plot*, nota-se os pontos aberrantes que foram observados anteriormente se distanciam consideravelmente da reta de referência, que contribuem para o desvio da normalidade dos resíduos quantílicos.

A fim de avaliar mais detalhadamente a adequação do modelo foi elaborado o *worm plot* dos resíduos quantílicos, apresentado na Figura 16

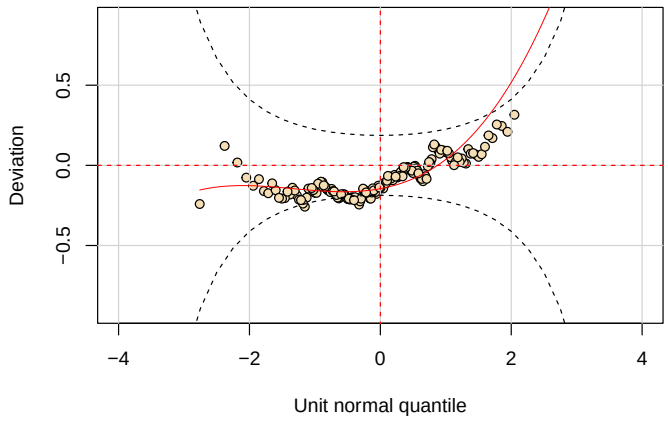


Figura 16: *Worm plot* dos resíduos quantílicos sob o Modelo 2 da Aplicação 2.

A partir da Figura 16, nota-se alguns pontos que se encontram fora das bandas de confiança, além de uma tendência de inclinação positiva dos pontos, o que sugere que o modelo não foi capaz de acomodar perfeitamente a sobredispersão dos dados. Neste aspecto, um modelo com distribuição que apresenta maior flexibilidade na variância pode acomodar melhor a sobredispersão. Vale destacar que a presença de pontos fora das bandas de confiança e a tendência observada permaneceram após o recálculo dos resíduos quantílicos.

Então, foi ajustado um GAMLSS sob distribuição binomial negativa, denotado por Modelo 3, a fim de se tentar acomodar a sobredispersão. Para o submodelo de μ, o parâmetro foi modelado em relação a ambas as covariáveis, com função de ligação logarítmica usual. Em relação ao submodelo de σ, o parâmetro de dispersão também foi modelado em função de ambas as covariáveis e utilizando função de ligação logarítmica.

A Figura 17 apresenta os gráficos de diagnóstico utilizando resíduos quantílicos a fim de avaliar a qualidade do ajuste do modelo. Pela Figura 17 (a), nota-se um leve aumento da variância dos resíduos quantílicos quando se aumentam os valores ajustados, porém, não apresenta desvio grave das suposições do modelo em relação à estrutura de variância. A Figura 17 (b) apresenta uma distribuição aleatória dos pontos ao redor de zero, ou seja, nenhum indício de correlação entre as observações. Além disso, não há resíduos com valores muito distantes de zero como nos modelos anteriores. A partir da Figura 17 (c), não se observa desvio grave da normalidade dos resíduos. Porém, a partir da Figura 17 (d), nota-se pontos se distanciando de forma considerável na cauda direita da reta de referência.

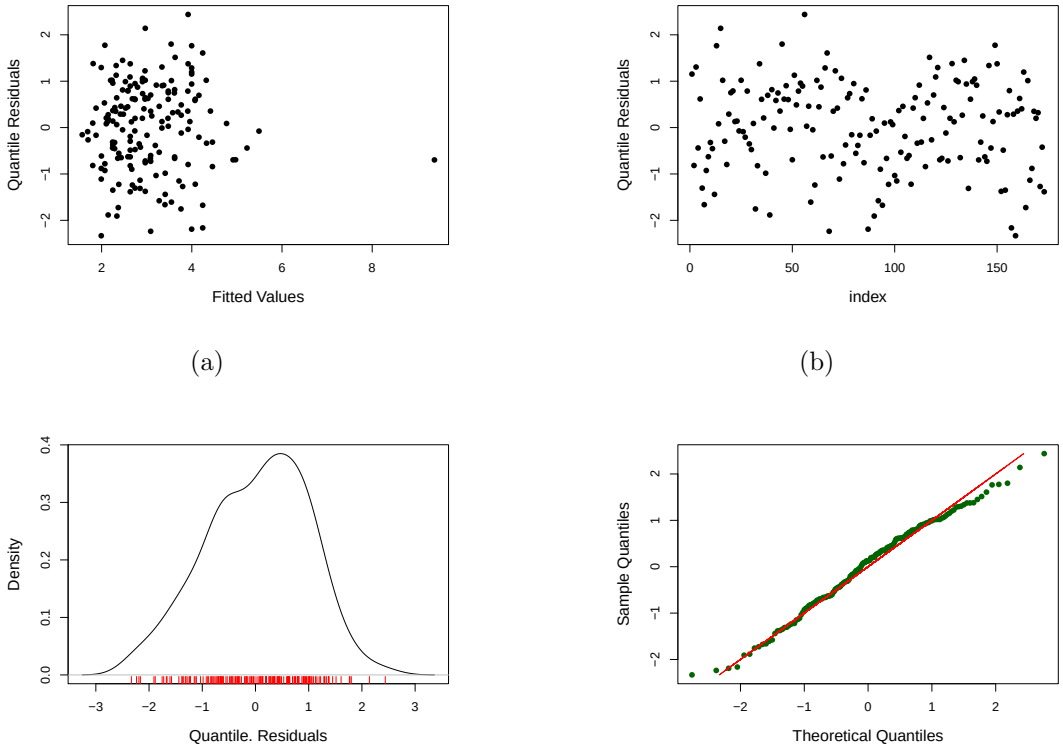


Figura 17: Gráficos de diagnóstico para os resíduos quantílicos sob o Modelo 3 da Aplicação 2: (a) dispersão dos resíduos pelos valores ajustados de μ; (b) dispersão dos resíduos pelos índices das observações; (c) densidade estimada dos resíduos; (d) *QQ-plot* dos resíduos.

A Figura 18 apresenta o *worm plot* dos resíduos quantílicos a fim de avaliar mais detalhadamente a adequação do ajuste do Modelo 3. Pelo gráfico, observa-se forte tendência dos pontos, o que compromete a qualidade do ajuste do modelo. Além disso, observou-se comportamento similar ao se recalcular os resíduos quantílicos. Como os

modelos que consideram apenas inflação de zeros (Modelo 2) ou apenas sobredispersão (Modelo 3) não foram capazes de acomodar adequadamente os dados, recomenda-se ajustar um modelo que incorpore simultaneamente ambas as características.

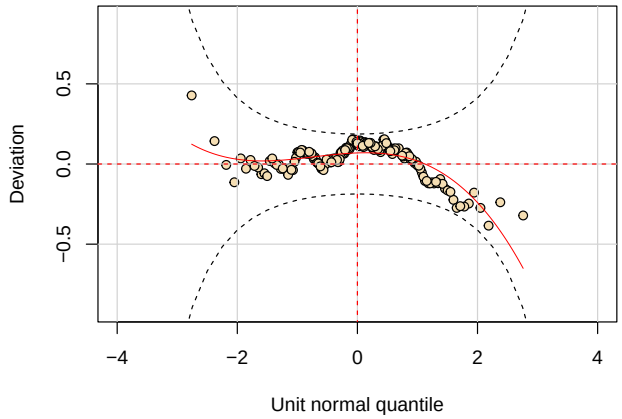


Figura 18: *Worm plot* dos resíduos quantílicos sob o Modelo 3 da Aplicação 2.

Dessa forma, um GAMLSS sob distribuição binomial negativa inflacionada em zero, denotado por Modelo 4, foi ajustado a fim de melhorar a adequação do ajuste. Analisaram-se diferentes arranjos entre covariáveis e funções de ligação empregados nos submodelos dos parâmetros, então optou-se por modelar o submodelo de μ em função de ambas as covariáveis usando, mais uma vez, função de ligação usual logarítmica. Para o submodelo de σ, considerou-se apenas o intercepto, pois as covariáveis não apresentaram significância para o parâmetro, sob função de ligação raiz quadrada, que contribuiu na melhora da qualidade do ajuste do modelo. Por fim, para modelar o parâmetro de inflação de zeros ν, utilizou-se a variável peso com função de ligação logito, por tratar-se de um parâmetro no intervalo unitário. Essa configuração foi escolhida por apresentar menos desvios em relação às demais.

Na Figura 19 são apresentados os gráficos de diagnóstico utilizando resíduos quantílicos a fim de avaliar a qualidade do ajuste do modelo. Pela Figura 19 (a), é perceptível uma leve heterogeneidade da variância resíduos quantílicos ao longo do gráfico, principalmente em razão dos pontos que apresentam resíduos um pouco mais elevados. A Figura 19 (b) apresenta uma distribuição aleatória dos pontos ao redor de zero, ou seja, nenhum indício de correlação entre as observações. A partir da Figuras 19 (c), percebe-se certo desvio da densidade esperada normal pelos resíduos quantílicos, com destaque para alguns resíduos de valor muito elevado. Por último, a Figura 19 (d) mostra desvio considerável de certa quantidade de ponto da reta de referência, o que é um indicativo de desvio da normalidade dos resíduos quantílicos.

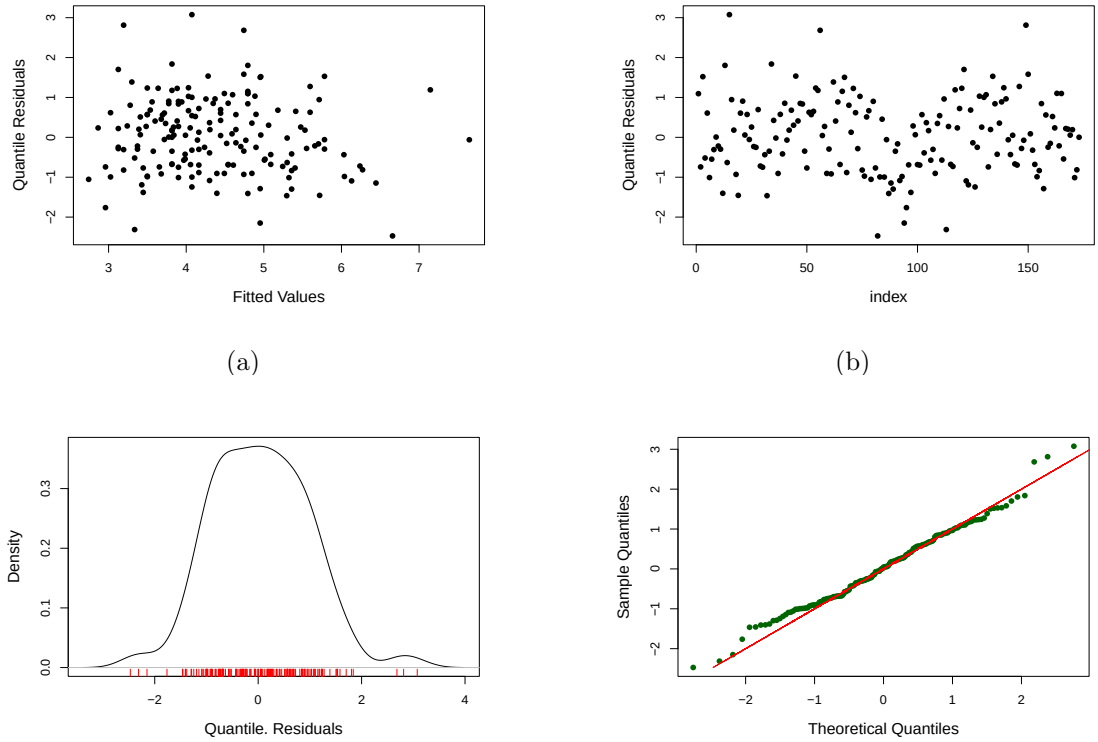


Figura 19: Gráficos de diagnóstico para os resíduos quantílicos sob o Modelo 4 da Aplicação 2: (a) dispersão dos resíduos pelos valores ajustados de μ; (b) dispersão dos resíduos pelos índices das observações; (c) densidade estimada dos resíduos; (d) *QQ-plot* dos resíduos.

Para aprofundar a análise diagnóstica do Modelo 4 foi elaborado um *worm plot*, apresentado na Figura 20.

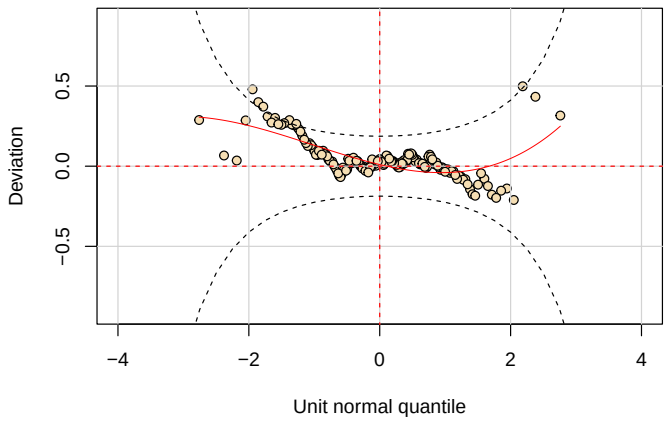


Figura 20: *Worm plot* dos resíduos quantílicos sob o Modelo 4 da Aplicação 2.

Pelo *worm plot* da Figura 20, é possível observar alguns pontos se fora das bandas de confiança. Adicionalmente, observa-se que os pontos não se concentram em torno de zero, sugerindo que o modelo não apresenta um ajuste satisfatório. Esse padrão se manteve mesmo após o recálculo dos resíduos quantílicos. Assim como ocorreu no Modelo 2, observou-se que alguns resíduos quantílicos ficaram moderadamente elevados devido a valores extremos da variável resposta, os quais interferem no bom ajuste do modelo. Nesse sentido, uma distribuição com caudas mais pesadas pode acomodar melhor essas observações, evitando a necessidade de removê-las.

Por fim, foi, então, ajustado o Modelo 5 sob distribuição Sichel inflacionada de zeros, que possui quatro parâmetros, e se caracteriza por ser uma distribuição que apresenta sobredispersão, caudas pesadas e inflação de zeros, a fim de se acomodar as observações que causaram problemas em modelos anteriores. Após a consideração de múltiplas combinações entre covariáveis e funções de ligação nos submodelos dos parâmetros, para μ, modelou-se em função de ambas as variáveis explicativas, por meio da função de ligação logarítmica. Em relação ao submodelo do parâmetro não interpretável σ, foi modelado em função da cor da carapaça utilizando função de ligação logarítmica, por se tratar também um parâmetro estritamente positivo. Sobre o parâmetro ν, também não interpretável, esse foi ajustado apenas com intercepto e usando função de ligação identidade, por se tratar de um parâmetro cujo domínio está nos reais. Por último, o parâmetro de inflação de zeros τ, foi ajustado em função da variável cor da carapaça, utilizando a função de ligação usual logito. Essa especificação foi escolhida por produzir desvios inferiores aos observados nas demais combinações.

A Figura 21 apresenta os diagnósticos elaborados a partir dos resíduos quantílicos, que permitem avaliar se o modelo representa adequadamente os dados. Na Figura 21 (a), há leve indicação de diminuição da variância dos resíduos quantílicos conforme os valores ajustados aumentam, porém, não representa um desvio grave das suposições do modelo. Na Figura 21 (b), os resíduos se dispersam de forma aleatória ao redor de zero, indicando não dependência entre as observações. Além disso, nota-se que as observações anteriormente associadas a resíduos excessivamente altos apresentaram redução considerável após o ajuste do modelo. Pelas as Figuras 21 (c) e 21 (d), percebe-se que os resíduos quantílicos apresentaram comportamento próximo ao da distribuição esperada normal, com a densidade estimada bem próxima do esperado e pontos seguindo a reta de referência no *QQ-plot*.

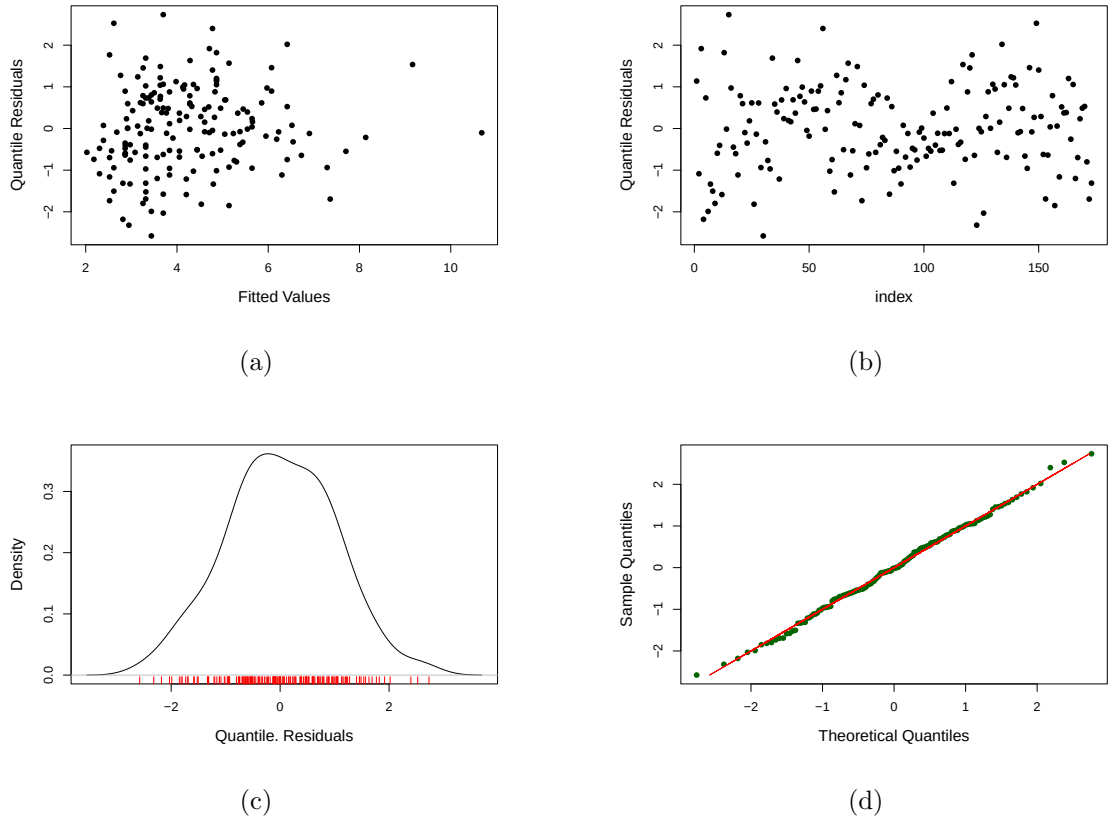


Figura 21: Gráficos de diagnóstico para os resíduos quantílicos sob o Modelo 5 da Aplicação 2: (a) dispersão dos resíduos pelos valores ajustados de μ; (b) dispersão dos resíduos pelos índices das observações; (c) densidade estimada dos resíduos; (d) *QQ-plot* dos resíduos.

Para finalizar a análise diagnóstica do Modelo 5 foi elaborado um*worm plot* dos resíduos quantílicos, apresentado na Figura 22.

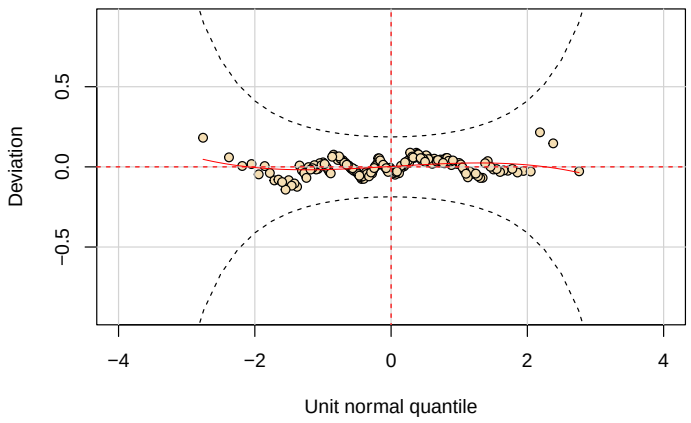


Figura 22: *Worm plot* dos resíduos quantílicos sob o Modelo 5 da Aplicação 2.

A partir da Figura 22 é possível observar que os pontos se localizam próximo zero e não apresentam qualquer tendência que indique inadequação do modelo. Vale destacar, ainda, que foi observado comportamento semelhante quando os resíduos quantílicos foram recalculados. Portanto, em razão da melhora em relação ao modelo anterior, serão apresentadas as estimativas e interpretações do Modelo 5.

A Tabela 7 apresenta as estimativas dos coeficientes, os erros-padrão, as estatísticas z do teste de Wald e p-valores dos teste de Wald sob o Modelo 5. No submodelo de μ, que representa a média condicional da distribuição desconsiderando os zeros estruturais, $\beta_{\mu 0}$ corresponde ao intercepto e $\beta_{\mu 1}$ ao coeficiente associado ao peso do caranguejo. Os parâmetros $\beta_{\mu 2_1}$, $\beta_{\mu 2_2}$ e $\beta_{\mu 2_3}$ referem-se, respectivamente, aos efeitos das categorias de coloração média, média escura e escura da carapaça, em que a coloração média clara é a referência. No submodelo de σ, cujos parâmetros não possuem interpretação direta em termos das contagens, $\beta_{\sigma 0}$ representa o intercepto, enquanto $\beta_{\sigma 2_1}$, $\beta_{\sigma 2_2}$ e $\beta_{\sigma 2_3}$ correspondem aos coeficientes associados às mesmas categorias de coloração consideradas em μ. O parâmetro β_ν representa o intercepto do submodelo de ν, também não interpretável diretamente. Por fim, no submodelo do parâmetro de inflação de zeros τ, $\beta_{\tau 0}$ representa o intercepto, e $\beta_{\tau 2_1}$, $\beta_{\tau 2_2}$ e $\beta_{\tau 2_3}$ representam os efeitos das categorias de coloração média, média escura e escura da carapaça do caranguejo-ferradura sobre a probabilidade de ocorrência de zeros estruturais.

Tabela 7: Estimativas, erros-padrão, estatísticas z e p-valores dos coeficientes sob o Modelo 5.

Coeficiente	Estimativa	Erro-Padrão	Estatística z	p-valor
$\beta_{\mu 0}$	0,7089	0,3893	1,821	0,069
$\beta_{\mu 1}$	0,3656	0,1321	2,768	0,006
$\beta_{\mu 2_1}$	-0,2415	0,1863	-1,296	0,195
$\beta_{\mu 2_2}$	-0,4416	0,2342	-1,885	0,059
$\beta_{\mu 2_3}$	0,3274	0,2337	1,401	0,161
$\beta_{\sigma 0}$	-3,0710	2,6770	-1,147	0,251
$\beta_{\sigma 2_1}$	1,6390	2,7370	0,599	0,549
$\beta_{\sigma 2_2}$	1,8660	2,8750	0,649	0,516
$\beta_{\sigma 2_3}$	-1,1890	6,4970	-0,183	0,855
β_ν	-1,9700	11,5700	-0,1700	0,865
$\beta_{\tau 0}$	-1,1393	0,6894	-1,653	0,098
$\beta_{\tau 2_1}$	-0,1051	0,7556	-0,139	0,889
$\beta_{\tau 2_2}$	0,3905	0,8285	0,471	0,637
$\beta_{\tau 2_3}$	1,8970	0,8280	2,291	0,022

A partir da Tabela 7, a um nível de significância de $\alpha = 0{,}10$, ambas as variáveis peso e coloração se mostraram significativas para explicar a média do número de satélites por fêmea, além disso, a variável da coloração da carapaça também foi significativa para explicar a inflação de zeros dos dados. Além disso, embora nenhuma das comparações com a categoria de referência no submodelo de σ tenha se mostrado significativa, essa configuração resultou no melhor ajuste do modelo. Vale mencionar, também, que, como pelo menos um dos níveis da coloração da carapaça apresentou coeficiente significativo em ambos os submodelos de μ e τ, a variável coloração da carapaça teve efeito significativo para os submodelos.

Como na aplicação anterior, em razão do uso da função de ligação logarítmica no submodelo de μ, as interpretações são feitas a partir da exponencial das estimativas. Nesse sentido, a estimativa do intercepto $\hat{\beta}_{\mu 0}$ não pode ser interpretada, pois, no contexto, não faz sentido fixar o peso da fêmea como zero. Em contrapartida, a estimativa $\hat{\beta}_{\mu 1}$ pode ser entendida por $\mathrm{e}^{\hat{\beta}_{\mu 1}} = \mathrm{e}^{0{,}3656} \approx 1{,}441$, ou seja, a adição de um quilograma ao peso da fêmea aumenta o número de satélites em 44,1% em média, desconsiderando os zeros estruturais. Em relação aos coeficientes estimados relativos a coloração da carapaça $\hat{\beta}_{\mu 2_1}$, $\hat{\beta}_{\mu 2_2}$ e $\hat{\beta}_{\mu 2_3}$, nota-se pelo sinal das estimativas que coloração média e média escura da carapaça impactam de forma negativa no número médio de satélites por caranguejo, enquanto a coloração escura tem efeito positivo. As exponenciais das estimativas com efeito negativo são 0,785 e 0,643, o que corresponde a reduções médias de 21,5% e 35,7%, respectivamente, considerando a média condicional sem os zeros estruturais. Por outro lado, a coloração escura apresenta efeito positivo, cuja exponencial resulta em 1,387, indicando um aumento médio de 38,7% no número de satélites em comparação à categoria de referência.

Por fim, a interpretação do intercepto associado ao submodelo de inflação de zeros pode ser obtida por meio da função inversa da logito. Assim, $\hat{\tau} = \mathrm{logit}^{-1}(\hat{\beta}_{\tau 0}) = \frac{1}{1+\mathrm{e}^{-\hat{\beta}_{\tau 0}}} = \frac{1}{1+\mathrm{e}^{1{,}1393}} \approx 0{,}199$ o que indica que, quando se trata de caranguejo de carapaça cor média clara, cerca de 19,9% das observações possuem probabilidade de gerar um zero estrutural, isto é, uma contagem nula de satélites que não provém da parte de contagem do modelo, ou seja, não provém da distribuição Sichel original, mas sim da inflação de zeros. Nesse sentido, esses zeros estruturais correspondem a situações em que a fêmea apresenta uma contagem nula de satélites não por variabilidade do processo de contagem, mas porque há algum fator que efetivamente impossibilita a ocorrência de machos satélites. Em relação à covariável da coloração da carapaça, a interpretação da razão de chances da inflação de zeros pode ser obtida por meio da exponencial das estimativas, considerando

a coloração média clara como referência. Nesse sentido, para caranguejos com coloração média, obtém-se $e^{\hat{\beta}_{\tau 2_1}} = e^{-0,1051} \approx 0,90$, indicando que a razão de chances de gerar um zero estrutural é cerca de 10% menor em relação à categoria de referência. Para a coloração média escura, a razão de chances é dada por $e^{\hat{\beta}_{\tau 2_2}} = e^{0,3905} \approx 1,48$, sugerindo uma razão de chances aproximadamente 48% maior de ocorrência de zeros estruturais quando comparada à referência. Por fim, para a coloração escura da carapaça, obtém-se $e^{\hat{\beta}_{\tau 2_3}} = e^{1,897} \approx 6,66$, o que indica que caranguejos-fêmeas com coloração escura apresentam uma razão de chances cerca de 6,66 vezes de gerar um zero estrutural em relação à coloração média clara. Ou seja, fêmeas com carapaça mais escura apresentam maior probabilidade estrutural de não possuírem satélites, isto é, a chance de realmente não ocorrer nenhum satélite aumenta conforme a carapaça se torna mais escura.

4 Conclusão

Este trabalho apresentou uma revisão abrangente dos principais modelos de regressão utilizados para dados de contagem, destacando suas características, limitações e aplicabilidade em diferentes cenários. Inicialmente, foram discutidos os modelos lineares generalizados, que representam um avanço importante em relação ao modelo linear normal, embora ainda imponham restrições relevantes, principalmente por modelarem apenas a média da distribuição e por assumirem estruturas que podem não refletir de forma adequada a variabilidade observada nos dados. Nesse contexto, os modelos aditivos generalizados para localização, escala, por modelarem até quatro parâmetros da distribuição, mostram-se como uma alternativa mais flexível e abrangente. Essa possibilidade amplia de maneira significativa a capacidade de ajuste e interpretação, especialmente em situações em que diferentes aspectos da distribuição dependem de covariáveis. Além disso, foram revisadas as principais distribuições discretas empregadas nesse tipo de modelagem. Também foram discutidos os tipos de resíduos mais utilizados no contexto dos MLGs e dos GAMLSSs, com destaque para os resíduos quantílicos, que se apresentam como uma ferramenta fundamental na avaliação do ajuste de modelos para dados de contagem. Além disso, foram descritas as principais técnicas gráficas empregadas na análise diagnóstica, as quais auxiliam na verificação da adequação dos modelos e na identificação de possíveis problemas de ajuste

As aplicações desenvolvidas ilustraram como diferentes distribuições são capazes de capturar padrões distintos presentes em dados reais. Além disso, foi apresentada de maneira detalhada a análise diagnóstica por meio dos resíduos quantílicos, que permite avaliar a adequação dos modelos e garantir resultados mais confiáveis e interpretáveis. Por meio de duas aplicações a conjuntos de dados reais, evidenciou-se a flexibilidade e a capacidade descritiva dos GAMLSSs. Na primeira aplicação, o modelo com distribuição binomial negativa, que permite modelar simultaneamente a média e o parâmetro de dispersão, mostrou-se adequado para capturar a sobredispersão dos dados e fornecer interpretações claras dos efeitos da idade e do tipo de tratamento sobre o número de lesões. Na segunda aplicação, a presença simultânea de sobredispersão, excesso de zeros e algumas observações com valores mais extremos exigiu o uso de modelos mais sofisticados, finalizando no ajuste de um modelo com distribuição Sichel inflacionada em zero, que incorpora quatro parâmetros e se mostrou capaz de modelar adequadamente a média da parte de contagem da distribuição e a inflação de zeros. Nessa aplicação, apesar da não significância da covariável coloração da carapaça no submodelo de σ, essa foi mantida por

contribuir substancialmente na melhoria da qualidade do ajuste, o que favorece a interpretação da variação dos parâmetros interpretáveis do modelo, considerando que σ não possui interpretação direta. Os resultados reforçam a importância da análise diagnóstica, com o uso de resíduos quantílicos e gráficos como o *worm plot*, para avaliar a adequação do modelo e nortear a escolha da distribuição mais apropriada.

Em síntese, os GAMLSSs representam uma ferramenta poderosa e versátil para a modelagem de dados de contagem, superando limitações de abordagens mais restritivas. Verifica-se também que a interpretação dos coeficientes e dos efeitos das covariáveis permanece acessível, o que facilita a utilização prática desses modelos. Dessa forma, este trabalho contribui para a disseminação das metodologias apresentadas ao demonstrar sua aplicação prática, desde a seleção da distribuição até a interpretação dos parâmetros estimados. Assim, espera-se que esse trabalho sirva como guia prático para análises futuras com dados em que se busque modelos capazes de acomodar padrões complexos e que ainda sejam interpretáveis.

Referências

BROCKMANN, H. J. Satellite male groups in horseshoe crabs, limulus polyphemus. *Ethology*, v. 102, n. 1, p. 1–21, 1996.

BUUREN, S. V.; FREDRIKS, M. Worm plot: a simple diagnostic device for modelling growth reference curves. *Statistics in Medicine*, v. 20, n. 8, p. 1259–1277, 2001.

CORDEIRO, G. M.; DEMÉTRIO, C. G. B. *Modelos Lineares Generalizados e Extensões*. Piracicaba, SP, Brasil: USP – ESALQ / Departamento de Ciências Exatas, 2008. Primeira edição.

COSTA, J.; SILVEIRA, L.; DONALISIO, M. Análise espacial de dados de contagem com excesso de zeros aplicado ao estudo da incidência de dengue em campinas, são paulo, brasil. *Cadernos de Saúde Pública*, v. 32, n. 8, p. e00036915, 2016.

COSTA, J. O.; JACQUES, M. A. P.; FREITAS, E. F.; PEREIRA, P. A. A. Acidentes rodoviários do tipo colisões nas estradas nacionais na região do minho: modelos de previsão para segmentos rodoviários. *Universidade do Minho*, 2012.

DUNN, P. K.; SMYTH, G. K. Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, v. 5, n. 3, p. 236–244, 1996.

DUNN, P. K.; SMYTH, G. K. *GLMsData: Generalized Linear Model Data Sets*. [S.l.], 2022. R package version 1.4. Disponível em: ⟨https://CRAN.R-project.org/package=GLMsData⟩.

GIARDIELLO, F. M.; HAMILTON, S. R.; KRUSH, A. J.; PIANTADOSI, S.; HYLIND, L. M.; CELANO, P.; BOOKER, S. V.; ROBINSON, C. R.; OFFERHAUS, G. J. A. Treatment of colonic and rectal adenomas with sulindac in familial adenomatous polyposis. *New England Journal of Medicine*, v. 328, n. 18, p. 1313–1316, 1993.

KUTNER, M. H.; NACHTSHEIM, C. J.; NETER, J.; LI, W. *Applied Linear Statistical Models 5th Edition*. [S.l.]: McGraw-Hill Irwin, 2005.

PAULA, G. A. *Modelos de Regressão: com Apoio Computacional*. São Paulo: Instituto de Matemática e Estatística, Universidade de São Paulo, 2025.

PEDRO, T. B.; SILVEIRA, A. K.; TASSINARI, W. d. S. Utilização de modelos de regressão aplicados a dados com inflacionamento de zeros: estudo da ocorrência de carrapatos em municípios no interior do estado do rio de janeiro. *Blucher Proceedings*, v. 1, n. 1, p. 601–612, 2014.

RIGBY, R. A.; STASINOPOULOS, D. M. Generalized additive models for location, scale and shape (with discussion). *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, Wiley, v. 54, n. 3, p. 507–554, 2005.

RIGBY, R. A.; STASINOPOULOS, M. D.; HELLER, G. Z.; BASTIANI, F. *Distributions for Modeling Location, Scale, and Shape: Using GAMLSS in R*. [S.l.]: Chapman and Hall/CRC, 2019.

STASINOPOULOS, D. M.; RIGBY, R. A.; HELLER, G. Z.; VOUDOURIS, V.; BASTIANI, F. D. *Flexible Regression and Smoothing: Using GAMLSS in R*. Boca Raton, Florida: Chapman and Hall/CRC, 2017.

STASINOPOULOS, D. M.; RIGBY, R. A.; VOUDOURIS, V.; AKANTZILIOTOU, C.; ENEA, M.; KIOSE, D.; ZEILEIS, A. *gamlss: Generalized Additive Models for Location Scale and Shape*. [S.l.], 2024. R package version 5.4-12. Disponível em: ⟨https://CRAN.R-project.org/package=gamlss⟩.