UNIVERSIDADE FEDERAL DO CEARÁ
CENTRO DE CIÊNCIAS
DEPARTAMENTO DE ESTATÍSTICA E MATEMÁTICA APLICADA
CURSO DE GRADUAÇÃO EM ESTATÍSTICA

JOSÉ DIEGO SOUZA LIRA

FERRAMENTAS DE DIAGNÓSTICO PARA DADOS DE CONTAGEM INFLACIONADOS DE ZEROS

FORTALEZA
2026

JOSÉ DIEGO SOUZA LIRA

FERRAMENTAS DE DIAGNÓSTICO PARA DADOS DE CONTAGEM INFLACIONADOS DE ZEROS

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Estatística do Centro de Ciências da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em Estatística.

Orientadora: Prof^{a}. Dra. Sílvia Maria de Freitas.

FORTALEZA

2026

JOSÉ DIEGO SOUZA LIRA

FERRAMENTAS DE DIAGNÓSTICO PARA DADOS DE CONTAGEM INFLACIONADOS DE ZEROS

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Estatística do Centro de Ciências da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em Estatística.

Aprovada em:

BANCA EXAMINADORA

Prof^{a}. Dra. Sílvia Maria de Freitas (Orientadora)
Universidade Federal do Ceará (UFC)

Prof. Dr. Juvêncio Santos Nobre
Universidade Federal do Ceará (UFC)

Prof. Dr. Gualberto Segundo Algamez Montalvo
Universidade Federal do Ceará (UFC)

À minha família Jacqueline e Apualpa.

AGRADECIMENTOS

Aos meus pais, Jacqueline e Apualpa, que me criaram com amor, dedicação e nunca mediram esforços para me proporcionar um futuro melhor. À minha madrinha e ao meu falecido padrinho, por jamais terem saído do meu lado e por sempre apoiarem minhas decisões ao longo de toda a minha vida.

À Profa. Dra. Sílvia Maria, por ter me acolhido como orientando durante os meus últimos dois anos de graduação, proporcionando-me a oportunidade de atuar como bolsista de monitoria e orientando minha monografia. Embora nunca tenha falado isso para ela, considero-a uma verdadeira mãe acadêmica, pelo apoio, cuidado e orientação recebidos ao longo da minha trajetória na UFC.

Um agradecimento especial ao Prof. Dr. Gualberto Segundo Agamez Montalvo, por ter sido meu professor em diversas disciplinas, por acreditar no meu potencial e por contribuir significativamente para a minha formação acadêmica. Além de ser um profissional extremamente competente e dedicado, é também uma pessoa admirável.

Ao Prof. Dr. João Maurício Araújo Mota, por ter ampliado minha visão sobre a Estatística e por me apresentar um mundo de possibilidades na área, até então ainda pouco claro para mim.

Ao Prof. Dr. Juvêncio Santos Nobre, que, embora tenha ministrado apenas uma disciplina para mim, rapidamente se tornou um dos meus professores preferidos, por quem nutro grande carinho, respeito e admiração.

Aos amigos e amigas que construí ao longo de toda a minha trajetória acadêmica, com quem compartilhei risadas, momentos, alegrias, conquistas e, algumas vezes, frustrações.

Por fim, expresso minha sincera gratidão à Universidade Federal do Ceará (UFC), por ter contribuído para o meu amadurecimento e desenvolvimento pessoal. Não vou esquecer todas as experiências e momentos vividos nesta instituição.

"Um sonho que você sonha sozinho é apenas um sonho. Um sonho que você sonha junto é realidade." (John Lennon)

RESUMO

Dados de contagem com excesso de zeros são frequentes em diversas áreas aplicadas, tais como saúde, ciências sociais e agronomia, apresentando desafios para modelagem tradicional baseada inicialmente na distribuição de Poisson, que pressupõe equidispersão. Neste trabalho, foram estudados e aplicados modelos da ordem dos Modelos Lineares Generalizados (MLGs) para lidar com tais conjuntos de dados, em particular o modelo Poisson, o modelo Binomial Negativo e os modelos inflacionados de zeros ZIP (Zero-Inflated Poisson) e ZINB (Zero-Inflated Negative Binomial). O objetivo central foi comparar o desempenho desses modelos em um conjunto de dados reais provenientes de um estudo sobre terapia conjugal, no qual a variável resposta é o número de passos em direção ao divórcio (MSI), caracterizada por excesso de zeros e superdispersão. Além da modelagem, foram desenvolvidas e aplicadas ferramentas de diagnóstico, como resíduos de Pearson, de *Deviance* e quantílicos, gráficos de HNP, Quantil-Quantil com envelopes simulados, *Worm-plot* e Distância de Cook, visando avaliar a qualidade do ajuste e identificar possíveis observações influentes. Os resultados indicaram que os modelos ZIP e ZINB apresentaram melhor adequação aos dados, com menor AIC e BIC, resíduos com melhor comportamento. O teste de Vuong não apontou diferença significativa entre ZIP e ZINB, sugerindo que o modelo ZIP, por ser mais parcimonioso, pode ser preferível para o conjunto analisado.

Palavras-chave: Dados de contagem; Superdispersão; Excesso de zeros; Modelos inflacionados de zeros; ZIP; ZINB; Diagnóstico de modelos.

ABSTRACT

Count data with excess zeros are common in several applied fields, such as health sciences, social sciences, and agronomy, posing challenges to traditional modeling approaches initially based on the Poisson distribution, which assumes equidispersion. In this study, models within the framework of Generalized Linear Models (GLMs) were investigated and applied to handle such data sets, namely the Poisson model, the Negative Binomial model, and the zero-inflated models ZIP (Zero-Inflated Poisson) and ZINB (Zero-Inflated Negative Binomial). The main objective was to compare the performance of these models using a real data set from a marital therapy study, in which the response variable is the number of steps toward divorce (MSI), characterized by excess zeros and overdispersion. In addition to model fitting, diagnostic tools were developed and applied, including Pearson and deviance residuals, quantile residuals, HNP plots, Quantile–Quantile plots with simulated envelopes, worm plots, and Cook's distance, in order to assess model adequacy and identify potential influential observations. The results indicated that the ZIP and ZINB models provided a better fit to the data, presenting lower AIC and BIC values and residuals with more satisfactory behavior. The Vuong test did not indicate a statistically significant difference between the ZIP and ZINB models, suggesting that the ZIP model, due to its greater parsimony, may be preferable for the analyzed data set.

Keywords: Data; Overdispersion; Excess of zeros; Zero-inflated models; ZIP; ZINB; Model diagnostics.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AFC	*Affective Communication*
AIC	Critério de Informação de Akaike
BIC	Critério de Informação Bayesiano
DAS	*Dyadic Adjustment Scale*
EMV	Estimação de Máxima Verossimilhança
FDA	Função de Distribuição Acumulada
HNP	*Half Normal Plot*
MLG	Modelo Linear Generalizado
MLGs	Modelos Lineares Generalizados
MSI	*Marital Status Inventory*
SEX	*Sexual Dissatisfaction*
ZINB	*Zero Inflated Negative Binomial*
ZIP	*Zero Inflated Poisson*

LISTA DE SÍMBOLOS

Y	Variável aleatória resposta
y_i	Valor observado da variável resposta para a i-ésima observação
μ	Média da variável resposta
n	Tamanho da amostra
p	Número parâmetros do modelo
β	Vetor de parâmetros de regressão
X	Matriz de especificação do modelo
x_i	Vetor de variáveis explicativas para a i-ésima observação
η	Preditor linear
$g(\cdot)$	Função de ligação
ϕ	Parâmetro de dispersão
θ	Vetor de parâmetros desconhecidos
$L(\theta)$	Função de Verossimilhança
$l(\theta)$	Logaritmo da função de Verossimilhança
$U(\theta)$	Função Escore
$K(\theta)$	Matriz de Informação de Fisher
$V(\mu)$	Função de variância
w_i	Pesos da matriz W no processo iterativo
H	Matriz de projeção
h_{ii}	i-ésimo elemento da diagonal principal da matriz H
D	*Deviance*
r_P	Resíduos de Pearson
r_D	Resíduos de Deviance
r_Q	Resíduos Quantílicos
π	Probabilidade de ocorrência de zeros estruturais
$\Phi(\cdot)$	Função de Distribuição Acumulada da Normal Padrão

SUMÁRIO

1 INTRODUÇÃO

Dados correspondem à informações obtidas por meio de registro, coleta ou levantamento sobre pessoas, eventos, experimentos, sensores, entrevistas, textos, etc. Permitem fundamentação, análise e embasamento na tomada de decisões, podendo ser classificados em sua totalidade como qualitativos (nominais ou ordinais) ou quantitativos (contínuos ou discretos) (Bussab; Morettin, 2009).

Na classe de dados quantitativos discretos, destacam-se os dados de contagem, porém é bem importante esclarecer o que se entende por dados de contagem. A palavra contagem usualmente é utilizada para se referir ao ato de enumerar unidades, itens ou registrar a quantidade de vezes que um determinado evento aconteceu, como por exemplo a quantidade de vezes que pacientes foram atendidos em um hospital em um determinado dia específico, a quantidade de acidentes ocorridos em uma rodovia durante o mês anterior. Para a Estatística, dados de contagem podem ser considerados como observações pertencentes a um subconjunto próprio dos inteiros não negativos $\{0, 1, 2, \dots\}$ (Hilbe, 2014).

Dessa forma, são inadequados à modelagem de dados que sejam considerados, por exemplo, seguindo uma distribuição Normal, como acontece nos modelos de regressão usuais. Os Modelos Lineares Generalizados (MLGs), propostos por Nelder e Wedderburn (1972), vieram como uma alternativa para modelagem de dados não provenientes de uma distribuição Normal, como é o caso dos dados de contagem, que têm a distribuição Poisson como premissa inicial de modelagem (McCullagh; Nelder, 1989).

Utilizando a distribuição Poisson como base para modelagem de dados de contagem, observamos algumas limitações intrínsecas a essa distribuição de probabilidade, como é o caso da variância ser igual à média. Esse fato traz diversas restrições para o desenvolvimento de estimativas confiáveis sobre dados de contagem. Na prática, é incomum encontrar dados de contagem que se adequem à propriedade de equidispersão da distribuição Poisson (Agresti, 2002).

A superdispersão é uma característica frequente em dados de contagem. É comum observar que a variância empírica dos dados seja superior à média, caracterizando a superdispersão. Já a equidispersão ocorre quando há equivalência entre a variância e a média, enquanto a subdispersão se verifica quando a variância é inferior à média (Dupuy, 2019). Assim, a distribuição Binomial Negativo torna-se uma alternativa viável à Poisson,

sendo uma opção quando nos deparamos com superdispersão e pode ser considerada uma extensão da Poisson, vista como uma mistura de Poisson e Gama (Hilbe, 2007).

A superdispersão (variância $>$ média) pode ser acomodada de diversas formas: modelos Quasi-Poisson (Wedderburn, 1974), modelos inflacionados de zeros (Lambert, 1992) e modelos mais complexos como os hierárquicos (Min; Agresti, 2005), dentre outras.

Porém, existem diversas razões para a presença de superdispersão no conjunto de dados. Uma delas, bastante comum, diz respeito ao excesso de zeros, normalmente causado por zeros estruturais inerentes ao estudo ou experimento em desenvolvimento. Com base nessa ideia Lambert (1992) apresentou os modelos ZIP e ZINB, posteriormente melhor definido por Ridout *et al.* (1998), com o objetivo de modelar esse excesso de zeros em conjunto com métodos de diagnóstico, tornando possível obter estimativas confiáveis e diagnóstico para dados com essas características.

Portanto, esta monografia tem como objetivo modelar dados reais de contagem inflacionados de zeros por meio da distribuição de Poisson (Poisson, 1837), da distribuição Binomial Negativo (Montmort, 1713) e dos modelos ZIP e ZINB, conforme propostos por Lambert (1992). Busca-se, assim, analisar e comparar o desempenho de cada abordagem, bem como avaliar sua eficiência na obtenção de estimativas confiáveis. Adicionalmente, foi desenvolvida uma análise de diagnóstico dos modelos, com foco em resíduos de Pearson, de *Deviance* e resíduos quantílicos, com o intuito de comparar seu desempenho em relação aos modelos considerados.

Diversos trabalhos recentes dedicaram-se ao estudo e à aplicação de modelos para dados de contagem inflacionados de zeros, destacando a relevância do tema. Fumes (2009) analisa dados de questionários de frequência alimentar aplicados a idosos de uma cidade do interior de São Paulo. Um segundo estudo, proposto por Costa *et al.* (2016) aborda o problema metodológico do excesso de zeros em dados espaciais de contagem de doenças em pequenas áreas. Carvalho (2019) trabalha na análise de dados de contagem provenientes de experimentos agronômicos. Portanto, esses estudos reforçam a necessidade de uma implementação consistente na análise de diagnósticos para modelos tão importantes no contexto atual.

Esta monografia está organizada da seguinte forma: o Capítulo 1 apresenta a introdução, na qual são expostas as ideias centrais e a motivação do trabalho. No Capítulo 2 são apresentados os MLGs, além dos resíduos utilizados e os gráficos que auxiliam na

interpretação desses resíduos. Já no Capítulo 3 são desenvolvidos os modelos para dados de contagem abordados na monografia, acompanhados das respectivas expressões dos resíduos para cada modelo. No Capítulo 4 é apresentada a aplicação dos métodos discutidos ao longo do trabalho. Por fim, o Capítulo 5 reúne as conclusões e considerações finais a respeito do trabalho. Essa organização visa facilitar a navegação pelo texto e contribuir para o estudo de leitores interessados no tema.

2 MODELO LINEAR GENERALIZADO

Ao lidar com o modelo de regressão linear, a distribuição Normal desempenha um papel central. Os procedimentos de inferência associados a esse modelo assumem que o erro segue uma distribuição Normal, com média zero e variância constante, e também que a distribuição condicional da variável resposta, dado o conjunto de variáveis explicativas, segue uma normal Normal. (Myers *et al.*, 2010). No entanto, em diversas situações, as suposições da regressão linear clássica não são adequadas, especialmente quando a variável resposta não é contínua, como ocorre em dados de contagem ou binários.

Diante dessas limitações, Nelder e Wedderburn (1972) propuseram os MLGs. A ideia central consiste em ampliar as possibilidades para a distribuição da variável resposta, permitindo que ela pertença à família exponencial linear de distribuições. Além disso, busca-se oferecer maior flexibilidade na forma funcional que relaciona a média da variável resposta aos preditores do modelo (Paula, 2013). Muitas das distribuições mais utilizadas podem ser agrupadas em uma categoria chamada família exponencial de distribuições. Fazem parte dessa família, por exemplo, as distribuições Normal, Binomial, Binomial Negativo, Gama, Poisson, Normal Inversa, Multinomial, Beta e Logarítmica, entre outras (Cordeiro; Demétrio, 2008).

2.1 Família Exponencial linear de Distribuições

Com a introdução de um parâmetro de dispersão $\phi > 0$, Nelder e Wedderburn (1972) ampliaram a formulação da família exponencial, permitindo incorporar novas distribuições no componente aleatório do modelo, denominada família exponencial linear de dispersão. Representada por

$$f(y_i; \theta_i, \phi) = \exp\left\{\phi^{-1}\left[y_i\theta_i - b(\theta_i)\right] + c(y_i; \phi)\right\}, \tag{2.1}$$

em que $b(\cdot)$ e $c(\cdot)$ são funções conhecidas. Quando ϕ é conhecido, a família de distribuições (2.1) é comparável à família exponencial na forma canônica (Cordeiro; Demétrio, 2008).

Pode-se demonstrar, sob condições usuais de regularidade, falaremos mais sobre essa condições no Apêndice A, conforme Paula (2013), que

$$\mathbb{E}\left\{\frac{\partial \log f(Y_i;\theta_i,\phi)}{\partial \theta_i}\right\} = 0$$

e

$$\mathbb{E}\left[\frac{\partial^2 \log f(Y_i;\theta_i,\phi)}{\partial \theta_i^2}\right] = -\mathbb{E}\left[\left\{\frac{\partial \log f(Y_i;\theta_i,\phi)}{\partial \theta_i}\right\}^2\right],$$

com base na expressão apresentada em (2.1), podemos determinar a média da variável aleatória Y. Inicialmente consideramos que para qualquer distribuição de probabilidade, a integral da função densidade sobre o suporte é igual a 1, ou seja

$$\int_y f(y_i;\theta_i;\phi)\ dy_i = 1,$$

substituindo pela expressão da família exponencial (2.1), temos

$$\int_y \exp\left\{\phi^{-1}\left[y_i\theta_i - b(\theta_i)\right] + c(y_i;\phi)\right\}\ dy_i = 1,$$

em seguida diferenciamos ambos os lados da equação em relação a θ_i e utilizando a Regra de Leibniz, assim obtemos

$$\int_y \frac{\partial}{\partial \theta_i} \exp\left\{\phi^{-1}[y_i\theta_i - b(\theta_i)] + c(y_i;\phi)\right\}\ dy_i = 0,$$

continuando o desenvolvimento

$$\Rightarrow \int_y \exp\left\{\phi^{-1}[y_i\theta_i - b(\theta_i)] + c(y_i;\phi)\right\} \cdot \frac{\partial}{\partial \theta_i}\left\{\phi^{-1}[y_i\theta_i - b(\theta_i)] + c(y_i;\phi)\right\}\ dy_i = 0$$

$$\Rightarrow \int_y \exp\left\{\phi^{-1}[y_i\theta_i - b(\theta_i)] + c(y_i;\phi)\right\} \cdot \left\{\phi^{-1}\left[\frac{\partial}{\partial \theta_i} y_i\theta_i - \frac{\partial}{\partial \theta_i} b(\theta_i)\right] + \frac{\partial}{\partial \theta_i} c(y_i;\phi)\right\}\ dy_i = 0$$

$$\Rightarrow \int_y \exp\left\{\phi^{-1}[y_i\theta_i - b(\theta_i)] + c(y_i;\phi)\right\} \cdot \left\{\phi^{-1}\left[y_i - b^{(1)}(\theta_i)\right]\right\}\ dy_i = 0$$

$$\Rightarrow \int_y f(y_i;\theta_i,\phi) \cdot \{[y_i - b'(\theta_i)]\}\ dy_i = 0$$

$$\Rightarrow \int_y f(y_i;\theta_i,\phi) y_i\ dy_i - b^{(1)}(\theta_i) \int_y f(y_i;\theta_i,\phi)\ dy_i = 0$$

$$\Rightarrow \mathbb{E}(Y_i) = b^{(1)}(\theta_i),$$

em que $b^{(1)}$ representa a derivada de primeira ordem de b, assim

$$E(Y_i) = \mu_i = b^{(1)}(\theta_i) \tag{2.2}$$

De maneira análoga à obtenção da média, o cálculo da variância aplica-se à derivada segunda para determinar o segundo momento, etapa fundamental para a composição da variância, dessa forma desenvolvendo a expressão abaixo temos que

$$\begin{aligned}
&\int_y \frac{\partial^2}{\partial \theta_i^2} \exp\left\{\phi^{-1}[y_i\theta_i - b(\theta_i)] + c(y_i;\phi)\right\}\ dy_i = 0 \\
\Rightarrow &\int_y \frac{\partial}{\partial \theta_i} \left\{f(y_i;\theta_i,\phi)\cdot[\phi^{-1}(y_i - b^{(1)}(\theta_i)]\right\}\ dy_i = 0 \\
\Rightarrow &\int_y f(y_i;\theta_i,\phi)^{(1)}\cdot[\phi^{-1}(y_i - b^{(1)}(\theta_i)] + f(y_i;\theta_i,\phi)\cdot[\phi^{-1}(y_i - b^{(1)}(\theta_i)]^{(1)}\ dy_i = 0 \\
\Rightarrow &\int_y f(y_i;\theta_i,\phi)\cdot[\phi^{-1}(y_i - b^{(1)}(\theta_i)]^2 + \int_y f(y_i;\theta_i,\phi)\cdot[-\phi^{-1}b^{(2)}(\theta_i)]\ dy_i = 0 \\
\Rightarrow &\mathbb{E}[[\phi^{-1}(Y_i - b^{(1)}(\theta_i)]^2] - \phi^{-1}b^{(2)}(\theta_i)\int_y f(y_i;\theta_i,\phi)\ dy_i = 0 \\
\Rightarrow &\mathbb{E}[[\phi^{-1}(Y_i - b^{(1)}(\theta_i)]^2] = \phi^{-1}b^{(2)}(\theta_i),
\end{aligned}$$

e trabalhando a parcela $\mathbb{E}[[\phi^{-1}(Y_i - b^{(1)}(\theta_i)]^2]$, concluímos que

$$\begin{aligned}
&\mathrm{Var}[\phi^{-1}(Y_i - b^{(1)}(\theta_i)] = \mathbb{E}[(\phi^{-1}(Y_i - b^{(1)}(\theta_i))^2] - \{\mathbb{E}[(\phi^{-1}(Y_i - b^{(1)}(\theta_i))]\}^2 \\
\Rightarrow &\mathbb{E}[(\phi^{-1}(Y_i - b^{(1)}(\theta_i))^2] = \phi^{-2}\mathrm{Var}(Y_i) + \phi^{-1}[\mathbb{E}(Y_i) - b^{(1)}(\theta_i)]^2 \\
\Rightarrow &\mathbb{E}[(\phi^{-1}(Y_i - b^{(1)}(\theta_i))^2] = \phi^{-2}\mathrm{Var}(Y_i) + \phi^{-1}[\mathbb{E}(Y_i) - b^{(1)}(\theta_i)]^2 \\
\Rightarrow &\mathbb{E}[(\phi^{-1}(Y_i - b^{(1)}(\theta_i))^2] = \phi^{-2}\mathrm{Var}(Y_i),
\end{aligned}$$

voltando para a expressão original

$$\begin{aligned}
&\Rightarrow \phi^{-2}\mathrm{Var}(Y_i) = \phi^{-1}b^{(2)}(\theta_i) \\
&\Rightarrow \phi^{-2}\mathrm{Var}(Y_i) = \phi^{-1}b^{(2)}(\theta_i) \\
&\Rightarrow \mathrm{Var}(Y_i) = \phi b^{(2)}(\theta_i)
\end{aligned}$$

dessa forma

$$\mathrm{Var}(Y_i) = \phi b^{(2)}(\theta_i) := \phi \mathrm{V}(\mu_i). \tag{2.3}$$

A função que relaciona o parâmetro canônico θ com a média é representada por $\theta = q(\mu_i)$ e a função da média com a variância é denotada por $b^{(2)}(\theta_i) = V(\mu_i) = d\mu_i/d\theta_i$. Denomina-se $V(\mu_i)$ função de variância dependendo apenas de μ. Assim, é possível expressar o parâmetro natural θ da seguinte maneira

$$\theta_i = \int \mathrm{V}^{-1}(\mu_i) d\mu_i = q(\mu_i).$$

A seguir apresentamos um conjunto de distribuições pertencentes à família exponencial linear apresentada em Cordeiro e Demétrio (2008). Assim Tabela 1, resume para algumas distribuições, o parâmetro de dispersão ϕ, o parâmetro canônico θ, a função cumulante $b(\theta)$ e o termo de normalização $c(y, \phi)$.

Tabela 1 – Algumas distribuições na família exponencial linear

Distribuição	ϕ	θ	$b(\theta)$	$c(y, \phi)$
Normal	σ^2	μ	$\dfrac{\theta^2}{2}$	$-\dfrac{1}{2}\left[\dfrac{y^2}{\sigma^2} + \log(2\pi\sigma^2)\right]$
Poisson	1	$\log(\mu)$	e^{θ}	$-\log(y!)$
Binomial	1	$\log\left(\dfrac{\mu}{m-\mu}\right)$	$m\log(1+e^{\theta})$	$\log\binom{m}{y}$
Binomial Negativo	1	$\log\left(\dfrac{\mu}{\mu+\phi}\right)$	$-\phi\log(1-e^{\theta})$	$\log\left[\dfrac{\Gamma(\phi+y)}{\Gamma(\phi)y!}\right]$
Gama	ν^{-1}	$-\dfrac{1}{\mu}$	$-\log(-\theta)$	$\nu\log(\nu y) - \log y - \log\Gamma(\nu)$
Normal Inversa	σ^2	$-\dfrac{1}{2\mu^2}$	$-(-2\theta)^{1/2}$	$-\dfrac{1}{2}\left[\log(2\pi\sigma^2 y^3) + \dfrac{1}{\sigma^2 y}\right]$

Neste trabalho, focaremos nas distribuições dessa família voltadas para dados de contagem, especificamente Poisson e Binomial Negativo, que serão detalhadas no Capítulo 3.

2.2 Definição do modelo linear generalizado

Os MLGs são aplicáveis quando existe uma variável aleatória Y relacionada a um conjunto de variáveis explicativas $x_1, \ldots, x_p$. Considerando uma amostra de n observações (y_i, x_i), em que $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\top$ representa o vetor das variáveis explicativas, de modo que a classe dos MLGs é constituído por três componentes fundamentais, segundo (Cordeiro; Demétrio, 2008), a saber:

Componente Aleatório: composto por um conjunto de variáveis aleatórias independentes $Y_1, Y_2, \ldots, Y_n$, oriundas de uma mesma distribuição pertencente à família de distribuições (2.1) com médias $\mu_1, \mu_2, \ldots, \mu_n$. A relevância da família apresentada na teoria dos MLGs reside no fato de que ela possibilita a modelagem de dados com diferentes características, como distribuições assimétricas, variáveis de natureza discreta ou contínua, e também aquelas restritas a um intervalo específico do conjunto dos reais, por exemplo, o intervalo $(0, 1)$.

Componente Sistemático: as variáveis explicativas, também denominadas covariáveis, entram no modelo por meio de uma combinação linear, dando origem ao i-ésimo preditor linear, definido por

$$\eta_i = \sum_{k=1}^{p} x_{ik}\, \beta_k = \mathbf{x}_i^\top \boldsymbol{\beta},$$

em que $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^\top$ representa o vetor de covariáveis associado à i-ésima observação e $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^\top$ é o vetor de parâmetros do modelo.

De forma matricial, o conjunto de preditores lineares pode ser expresso como

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

sendo $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)^\top$ a matriz de especificação do modelo, de dimensão $n \times p$, e $\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_n)^\top$ o vetor de preditores lineares com dimensão $n \times 1$.

Função de ligação: a função que associa o componente aleatório ao componente sistemático, ligando a média ao seu preditor linear, representado por

$$\eta_i = g(\mu_i),$$

também expressa por

$$\mu_i = g^{-1}(\eta_i),$$

como a função $g(\cdot)$ sendo monótona e ao menos duplamente diferenciável em todo o seu domínio, garantindo a existência da inversa e a relação entre a média e o preditor linear.

2.3 Estimação por Máxima Verossimilhança

Considerando $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \phi)^\top$ e uma amostra aleatória independente $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^\top$, com $i = 1, 2, \ldots, n$, pertencente à família exponencial linear, as estimativas dos parâmetros $\boldsymbol{\beta}$ e ϕ podem ser obtidas pelo método de máxima verossimilhança, por meio da maximização do logaritmo da função de verossimilhança. Assim, a partir da função densidade de probabilidade dada em (2.1), a função de verossimilhança associada à $\boldsymbol{\theta}$ é expressa por

$$\begin{aligned} L(\boldsymbol{\theta}, \mathbf{y}) &= \prod_{i=1}^{n} \exp\left\{\phi^{-1}\left[y_i\theta_i - b(\theta_i)\right] + c(y_i;\phi)\right\} \\ &= \exp\left\{\phi^{-1}\sum_{i=1}^{n}\left[y_i\theta_i - b(\theta_i)\right] + \sum_{i=1}^{n} c(y_i;\phi)\right\}. \end{aligned}$$

em consequência disso, o logaritmo da função de verossimilhança é dado por

$$\begin{aligned} l(\boldsymbol{\theta}) &= \log(L(\boldsymbol{\theta}, \mathbf{y})) \\ &= \log\left\{\exp\left\{\phi^{-1}\sum_{i=1}^{n}\left[y_i\theta_i - b(\theta_i)\right] + \sum_{i=1}^{n} c(y_i;\phi)\right\}\right\} \\ &= \phi^{-1}\sum_{i=1}^{n}\left[y_i\theta_i - b(\theta_i)\right] + \sum_{i=1}^{n} c(y_i;\phi). \end{aligned}$$

Primeiramente, vamos obter a função escore associada à j-ésima componente do vetor $\boldsymbol{\beta}$, derivando-se $l(\boldsymbol{\theta}, \mathbf{y})$ em relação a cada coeficiente

$$\begin{aligned} U_{\beta_j}(\boldsymbol{\theta}) = \frac{\partial l(\theta)}{\partial \beta_j} &= \frac{\partial}{\partial \beta_j}\left\{\phi^{-1}\sum_{i=1}^{n}\left[y_i\theta_i - b(\theta_i)\right] + \sum_{i=1}^{n} c(y_i,\phi)\right\} \\ &= \phi^{-1}\sum_{i=1}^{n}\left[y_i\frac{\partial\theta_i}{\partial\beta_j} - \frac{\partial b(\theta_i)}{\partial\beta_j}\right] + \sum_{i=1}^{n}\frac{\partial c(y_i,\phi)}{\partial\beta_j} \\ &= \phi^{-1}\sum_{i=1}^{n}\left[y_i\frac{\partial\theta_i}{\partial\beta_j} - \frac{\partial b(\theta_i)}{\partial\beta_j}\right]. \end{aligned}$$

Conforme mostrado anteriormente, θ_i depende funcionalmente de μ_i, e η_i, correspondendo ao i-ésimo componente do vetor de preditores lineares. Além disso, μ_i é obtido a partir de η_i por meio da inversa da função de ligação. Considerando $j = 1, 2, \ldots, p$, em que p é o número de parâmetros do modelo, tem-se $\mu_i = b^{(2)}(\theta_i)$.

$$\mathrm{V}_i = \mathrm{V}(\mu_i) = \frac{d\mu_i}{d\theta_i} \Rightarrow \mathrm{V}_i^{-1} = \frac{d\theta_i}{d\mu_i},$$

então

$$\begin{aligned}
\frac{\partial l(\theta)}{\partial \beta_j} &= \phi^{-1} \sum_{i=1}^{n} \left(y_i \frac{d\theta_i}{d\mu_i} \frac{d\mu_i}{d\eta_i} \frac{d\eta_i}{d\beta_j} - \frac{db(\theta_i)}{d\theta_i} \frac{d\theta_i}{d\mu_i} \frac{d\mu_i}{d\eta_i} \frac{d\eta_i}{d\beta_j} \right) \\
&= \phi^{-1} \sum_{i=1}^{n} \left(y_i \mathrm{V}_i^{-1} \frac{d\mu_i}{d\eta_i} \frac{d\eta_i}{d\beta_j} - \frac{db(\theta_i)}{d\theta_i} \mathrm{V}_i^{-1} \frac{d\mu_i}{d\eta_i} \frac{d\eta_i}{d\beta_j} \right) \\
&= \phi^{-1} \sum_{i=1}^{n} \left(y_i \mathrm{V}_i^{-1} \frac{d\mu_i}{d\eta_i} x_{ij} - \frac{db(\theta_i)}{d\theta_i} \mathrm{V}_i^{-1} \frac{d\mu_i}{d\eta_i} \frac{d\eta_i}{d\beta_j} \right) \\
&= \phi^{-1} \sum_{i=1}^{n} \left(y_i \mathrm{V}_i^{-1} \frac{d\mu_i}{d\eta_i} x_{ij} - \mu_i \mathrm{V}_i^{-1} \frac{d\mu_i}{d\eta_i} x_{ij} \right) \\
&= \phi^{-1} \sum_{i=1}^{n} \left(\mathrm{V}_i^{-1} \frac{d\mu_i}{d\eta_i} (y_i - \mu_i) x_{ij} \right) \\
&= \phi^{-1} \sum_{i=1}^{n} \left(\sqrt{\frac{w_i}{\mathrm{V}_i}} (y_i - \mu_i) x_{ij} \right).
\end{aligned}$$

em que $w_i = \left(\frac{d\mu_i}{d\eta_i}\right)^2 \mathrm{V}_i^{-1}$. Assim, é possível reescrever a função escore na forma matricial

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

sendo $\mathbf{X}$ uma matriz especificação do modelo de dimensão $n \times p$, com linhas denotadas por $\mathbf{x}_i^\top$, $i = 1, 2, \ldots, n$; $\mathbf{W} = \mathrm{diag}\{w_1, w_2, \ldots, w_n\}$ conhecida como matriz de pesos; $\mathbf{V} = \mathrm{diag}\{\mathrm{V}_1, \mathrm{V}_2, \ldots, \mathrm{V}_n\}$, $\mathbf{y} = (y_1, y_2, \ldots, y_n)^\top$ e $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_n)^\top$.

Obtemos a matriz de informação de Fisher para $\boldsymbol{\beta}$ derivando uma segunda vez a $L(\boldsymbol{\theta})$ com relação aos coeficientes

$$K_{\beta_j\beta_l}(\boldsymbol{\theta}) = -\mathbb{E}\left[\frac{\partial^2 l(\theta)}{\partial\beta_j\partial\beta_l}\right] = \mathbb{E}\left[\frac{\partial l(\theta)}{\partial\beta_j}\frac{\partial l(\theta)}{\partial\beta_l}\right]$$
$$= \sum_{i=1}^{n}\mathbb{E}\left[\phi^{-1}V_i^{-1}\frac{\partial\mu_i}{\partial\eta_i}(y_i-\mu_i)\,x_{ij}\times\phi^{-1}V_i^{-1}\frac{\partial\mu_i}{\partial\eta_i}(y_i-\mu_i)\,x_{il}\right]$$
$$= \sum_{i=1}^{n}\phi^{-2}V_i^{-2}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2\mathbb{E}\big[(y_i-\mu_i)^2\big]\,x_{ij}x_{il},$$

considerando $Var(Y_i) = \mathbb{E}\big[(y_i-\mu_i)^2\big]$ e novamente $w_i = \left(\frac{d\mu_i}{d\eta_i}\right)^2 V_i^{-1}$, temos

$$K_{\beta_j\beta_l}(\boldsymbol{\theta}) = \sum_{i=1}^{n}\phi^{-2}V_i^{-1}w_i Var(Y_i)\,x_{ij}x_{il}.$$

Utilizando de forma apropriada a definição (2.3), temos

$$K_{\beta_j\beta_l}(\boldsymbol{\theta}) = \sum_{i=1}^{n}\phi^{-2}V_i^{-1}w_i\phi V_i\,x_{ij}x_{il}$$
$$= \phi^{-1}\sum_{i=1}^{n}w_i x_{ij}x_{il},$$

logo, a informação de Fisher é expressa na forma matricial por

$$K_{\beta\beta}(\boldsymbol{\theta}) = -\mathbb{E}\left[\frac{\partial^2 L(\boldsymbol{\theta})}{\partial\beta_j\partial\beta_l}\right] = \phi^{-1}\mathbf{X}^\top\mathbf{W}\mathbf{X}.$$

Já para a função escore do parâmetro ϕ a expressão fica dada por

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

logo, a informação de Fisher de ϕ é obtida da seguinte forma:

$$
\begin{aligned}
K_{\phi\phi}(\boldsymbol{\theta}) &= -\mathbb{E}\left[\frac{\partial^2 l(\theta)}{\partial \phi^2}\right] \\
&= -\mathbb{E}\left\{\frac{\partial^2}{\partial \phi^2}\left[\sum_{i=1}^{n}\left[\phi^{-1}\left[y_i\theta_i - b(\theta_i)\right] + c(y_i;\phi)\right]\right]\right\} \\
&= -\sum_{i=1}^{n}\mathbb{E}\left\{\left[\left[y_i\theta_i - b(\theta_i)\right]\frac{\partial}{\partial \phi^2}\phi^{-1} + \frac{\partial}{\partial \phi^2}c(y_i;\phi)\right]\right\},
\end{aligned}
$$

sendo o vetor de parâmetros $\boldsymbol{\beta}$ e ϕ ortogonais. Dessa forma, a matriz de informação de Fisher para $\boldsymbol{\theta}$ é dada pela matriz bloco diagonal:

$$
\mathbf{K}_{\theta\theta}(\boldsymbol{\theta}) = \operatorname{diag}\left\{K_{\beta\beta}(\boldsymbol{\theta}),\, K_{\phi\phi}(\boldsymbol{\theta})\right\} = \begin{bmatrix} K_{\beta\beta} & \mathbf{0} \\ \mathbf{0} & K_{\phi\phi} \end{bmatrix},
$$

e função escore dada por

$$
\mathbf{U}_{\theta} = \left(U_{\beta}^{\top},\, U_{\phi}\right)^{\top}.
$$

Nesse contexto, os estimadores Estimação de Máxima Verossimilhança (EMV) dos parâmetros $\boldsymbol{\beta}$ e $\boldsymbol{\phi}$, representados por $\hat{\boldsymbol{\beta}}$ e $\hat{\boldsymbol{\phi}}$, são obtidas igualando-se U_{β_j} e U_{ϕ} a zero, respectivamente em geral essas equações não são lineares e têm que ser resolvidas por métodos iterativos, como por exemplo o Newton-Raphson (Cordeiro; Demétrio, 2008).

2.4 *Deviance*

Ao desenvolver um estudo envolvendo o ajuste de umModelo Linear Generalizado (MLG), um passo fundamental consiste na avaliação da qualidade do modelo obtido. Nesse contexto, a *Deviance*, introduzida por Nelder e Wedderburn (1972), consolidou-se como uma das principais medidas para diagnóstico e comparação de modelos lineares generalizados encaixados. Essa medida permite quantificar o grau de discrepância entre o modelo ajustado e o modelo saturado.

Considerando um modelo mais simples nomeado como modelo nulo, com apenas um parâmetro que representa a média μ partilhado por todas as observações y's, e um modelo saturado com $(p = n)$ com o número de parâmetros igual ao número de observações

na amostra, toda a variação dos y's é alocada ao componente sistemático, seja função $L(\boldsymbol{\mu}; \mathbf{y})$ é expressa por

$$L(\mathbf{y}; \mathbf{y}) = \sum_{i=1}^{n} L(y_i; y_i).$$

Assim, a estimativa de máxima verossimilhança de μ_i, nesse caso, é simplesmente o próprio valor observado, $\tilde{\mu} = y_i$ (Cordeiro; Neto, 2006).

Na prática, o modelo nulo apresenta uma estrutura simples, enquanto o modelo saturado, não oferece informação útil para inferência. Contudo, esse modelo é importante porque permite quantificar a distância de um modelo intermediário com $(p < n)$ parâmetros em relação a um modelo considerado perfeito.

Considerando o logaritmo da função de verossimilhança definido por:

$$L(\boldsymbol{\mu}; \mathbf{y}) = \sum_{i=1}^{n} L(\mu_i; y_i),$$

com $\mu_i = g^{-1}(\eta_i)$ e $\eta_i = \mathbf{x}_i^{\top}\boldsymbol{\beta}$, e denotando a estimativa de $L(\boldsymbol{\mu}; \mathbf{y})$ por $L(\hat{\boldsymbol{\mu}}; \mathbf{y})$, dessa forma, a função desvio, que é a medida de qualidade de ajuste do modelo em estudo, é representada por

$$D^*(\mathbf{y}; \hat{\boldsymbol{\mu}}) = \phi D(\mathbf{y}; \hat{\boldsymbol{\mu}}) = 2[L(\mathbf{y}; \mathbf{y}) - L(\hat{\boldsymbol{\mu}}; \mathbf{y})], \tag{2.4}$$

representando a diferença entre o máximo da função log-verossimilhança entre o modelo saturado e o modelo em estudo avaliado na estimativa máxima verossimilhança de $\hat{\boldsymbol{\beta}}$, tal que $D^*(\mathbf{y}; \hat{\boldsymbol{\mu}}) = \sum_{i=1}^{n} d_i^2$ (Paula, 2013).

2.5 Análise de Resíduos

Considere um modelo geral no qual as observações $Y_1, \ldots, Y_n$ seguem

$$Y_i = w_i(\boldsymbol{\beta}, \epsilon_i),$$

com $i = 1, 2, \ldots, n$, em que g_i é uma função que relaciona o vetor de parâmetros $\boldsymbol{\beta}$ ao termo de erro ϵ_i, sendo este uma variável aleatória não observada. Dessa forma, podem-se definir os resíduos como

$$Y_i = w_i(\hat{\beta}_i, R_i),$$

podendo ser expresso da seguinte forma

$$R_i = h_i(Y_i, \hat{\beta}_i),$$

sendo $\hat{\beta}_i$ o estimador máxima verossimilhança do parâmetro β_i, R_i os resíduos brutos e h_i uma função inversa de g_i em relação ao ϵ_i, representada por

$$\epsilon_i = h_i(Y_i, \hat{\beta}_i).$$

Dessa forma, os resíduos R_i são uma predição dos erros aleatórios, quando utilizamos os EMV para estimar os parâmetros (Cox; Snell, 1968).

Na modelagem estatística, o processo de diagnóstico se destaca como a etapa muito importante para a escolha do melhor modelo, pois permite verificar desvios das suposições feitas para o modelo. No contexto dos MLGs os resíduos desempenham o papel de avaliar o modelo ajustado com relação à escolha da função de variância, da função de ligação e dos termos do preditor linear apropriado. Também, corroborando na avaliação da identificação de pontos aberrantes, que podem influenciar no ajuste do modelo em estudo (Cordeiro; Neto, 2006).

2.5.1 Resíduos de Pearson

Os resíduos de Pearson são definidos por

$$r_{P_i} = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}}, \tag{2.5}$$

resultantes da padronização dos resíduos ordinários pela raiz da função de variância. Recebem esse nome em razão de sua conexão com a estatística qui-quadrado de Pearson (χ^2), uma vez que, no caso da distribuição Poisson, o quadrado de r_{P_i} corresponde diretamente ao termo da estatística de Pearson (McCullagh; Nelder, 1989).

2.5.2 Resíduos de Pearson Padronizados

Definimos os resíduos de Pearson em sua forma padronizada por

$$r_{P_i}^{*} = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)\,(1 - \hat{h}_{ii})}}, \tag{2.6}$$

em que $\hat{h}_{ii}$, denominado *leverage*, mede o grau de influência da i-ésima observação na estimação dos parâmetros do modelo. O termo $\hat{h}_{ii}$ corresponde ao i-ésimo elemento da diagonal da matriz $\hat{\mathbf{H}}$, conhecida como matriz de projeção ou matriz chapéu, definida por

$$\hat{\mathbf{H}} = \hat{\mathbf{W}}^{1/2}\mathbf{X}\left(\mathbf{X}^{\top}\hat{\mathbf{W}}\mathbf{X}\right)^{-1}\mathbf{X}^{\top}\hat{\mathbf{W}}^{1/2}, \tag{2.7}$$

em que $\hat{\mathbf{W}}$ é a matriz de pesos e $\mathbf{X}$ é a matriz de especificação do modelo. A matriz $\hat{\mathbf{H}}$ é simétrica e idempotente, e satisfaz $\mathrm{tr}(\hat{\mathbf{H}}) = p$, em que p é o número de parâmetros do modelo. Além disso, seus elementos diagonais obedecem à propriedade $0 \leq h_{ii} \leq 1$.

2.5.3 Resíduos de Deviance

Utilizando as componentes d_i que formam a *Deviance*, apresentada em (2.4), os resíduos *Deviance* pode ser definido pela expressão

$$r_{D_i} = sinal(y_i - \hat{\mu}_i)\sqrt{d_i} \tag{2.8}$$

em que a transformação aplicada tem como objetivo aproximar a distribuição dos resíduos de uma distribuição Normal. A raiz quadrada das partes do desvio d_i gera resíduos que preservam as propriedades da transformação, gerando resíduos aproximadamente simétricos. Dessa forma, os resíduos r_{D_i} podem ser encarados como uma aproximação da normal padrão. Consequentemente, $r_{D_i}^2$ tem uma distribuição próxima $\chi_{(1)}^2$ (Cordeiro; Neto, 2006).

2.5.4 Resíduos de Deviance Padronizados

Assim como ocorre com os resíduos de Pearson, os resíduos de *Deviance* também possuem uma forma padronizada, levando em consideração o *leverage* em suas estimativas,

representada por

$$r_{D_i^*} = \frac{r_{D_i}}{\sqrt{(1-\hat{h}_{ii})}} \tag{2.9}$$

2.5.5 Resíduos Quantílicos Aleatorizados

Desenvolvido por Dunn e Smyth (1996), os resíduos quantílicos utilizam a inversa da Função de Distribuição Acumulada (FDA) do modelo ajustado com o objetivo de associar cada observação em um quantil da Normal padrão. Dessa forma, transformando os resíduos observados em valores que seguem, aproximadamente, uma distribuição Normal padrão, independente da distribuição da variável resposta. Permite diagnósticos mais precisos para variáveis aleatórias discretas ou em modelos com forte assimetria.

Para variáveis aleatórias contínuas, como no caso da Normal, Normal Inversa, Gama, Lognormal, etc, os resíduos quantílicos são expressos por

$$r_{q_i} = \Phi^{-1}[F(y_i; \hat{\mu}_i; \hat{\phi}_i)].$$

No caso de variáveis discretas, como para distribuições Poisson, Binomial, Binomial Negativo, etc, a FDA apresenta descontinuidades, o que implica que $F(y_i; \hat{\mu}_i, \hat{\phi}_i)$ não varia continuamente, não gerando resíduos contínuos. Para contornar essa limitação, é incluída uma aleatorização por meio de uma variável uniforme nos intervalos da FDA, resultando em

$$r_{q_i} = \Phi^{-1}\big(F(y_i^-; \hat{\mu}_i, \phi) + u_i \left[F(y_i; \hat{\mu}_i, \phi) - F(y_i^-; \hat{\mu}_i, \phi)\right]\big),$$

sendo Φ^{-1} a função quantil da Normal padrão inversa, $F(y_i; \hat{\mu}_i; \hat{\phi}_i)$ a função de distribuição acumulada de Y_i, $F(y_i^-; \hat{\mu}_i, \phi)$ representa a FDA imediatamente anterior ao termo Y_i e u_i uma variável aleatória uniforme no intervalo $(0, 1]$.

2.6 Estatísticas para Diagnóstico

Pontos atípicos podem ser caracterizados por apresentarem alta alavancagem (*leverage*) ou resíduos elevados, sendo também denominados observações inconsistentes

(McCullagh; Nelder, 1989). Observações inconsistentes distinguem-se por se afastarem do padrão observado nas demais observações, passando a serem consideradas influentes quando sua remoção do conjunto de dados resulta em alterações relevantes nas estimativas dos parâmetros e no ajuste global do modelo (Cordeiro; Demétrio, 2008).

2.6.1 Distância de Cook

Entre as medidas mais utilizadas para detectar tais observações influentes destaca-se a Distância de Cook, apresentada por Cook (1977). Essa medida combina resíduos padronizados com a medida de alavancagem, tornando-se um indicador global da influência que uma determinada observação apresenta durante o processo de ajuste do modelo. Assim, a influência da i-ésima observação sobre o vetor de parâmetros é expressa, em sua forma matricial, por

$$D_i = \frac{(\hat{\beta}_{(i)} - \hat{\beta})^\top X^\top X (\hat{\beta}_{(i)} - \hat{\beta})}{p\hat{\sigma}^2}$$

na qual, D_i representa a Distância de Cook para a observação i, $\hat{\beta}$ representa o vetor com todos os coeficientes do modelo original e $\hat{\beta}_{(i)}$, é o vetor de coeficientes sem a observação i, X a matriz de especificação do modelo, p a quantidade de parâmetros e $\hat{\sigma}^2$ a estimativa da variância dos erros.

A regra geral habitualmente empregada para determinar pontos influentes considera observações maiores com $D_i > 1$ altamente influentes, já observações com $D_i > 0,5$ merecem investigação. Porém, uma classificação mais rigorosa, considerada para amostras grandes, utiliza-se uma fórmula ajustada para o número de observações, definida por $4/n$, em que n representa o tamanho da amostra (Cordeiro; Neto, 2006).

2.7 Análise Gráfica dos Resíduos

2.7.1 Gráfico Índices

Apesar de simples, os gráficos de índices têm uma finalidade importante pois torna possível avaliar a estrutura dos resíduos ao longo das observações, permitindo identificar facilmente pontos influentes que podem prejudicar as estimativas.

2.7.2 Gráfico Resíduos versus Valores Ajustados

Nesse gráfico, os resíduos obtidos no ajuste são dispostos no eixo horizontal, enquanto os valores ajustados pelo modelo são apresentados no eixo vertical, permitindo a avaliação da linearidade e da relação entre as variáveis. Espera-se que os resíduos se distribuam aleatoriamente em torno de zero, indicando um bom ajuste da função de ligação. Segundo McCullagh e Nelder (1989), esse gráfico também contribui para a verificação da constância da variância da distribuição assumida, estando, portanto, relacionado à estrutura de variância dos modelos lineares generalizados .

2.7.3 Gráfico Half Normal Plot (HNP)

O gráfico de HNP baseia-se na comparação entre as estatísticas de ordem observadas, obtidas a partir do modelo ajustado, e as estatísticas de ordem esperadas sob uma distribuição meio-normal. Estas últimas são calculadas com elevada precisão por

$$\Phi^{-1}\left(\frac{i-\frac{3}{8}}{n+\frac{1}{4}}\right),$$

em que i representa a i-ésima estatística de ordem e n é o tamanho da amostra (McCullagh; Nelder, 1989).

2.7.4 Gráfico Quantis-Quantis com envelopes simulados

O *Q-Q plot* tem como finalidade comparar os quantis dos resíduos observados com os quantis teóricos da distribuição de referência. Quando os resíduos apresentam normalidade aproximada, espera-se que os pontos se alinhem ao longo de uma reta de 45°. Especificamente no contexto dos MLGs, utiliza-se a distribuição normal como referência. O acréscimo de envelopes simulados, que constituem intervalos de aceitação para a suposição de normalidade, é obtido por meio de simulações de Monte Carlo, contribuindo para uma avaliação visual mais robusta (Marden, 2004).

2.7.5 Gráfico Worm-Plot

Apresentado por Buuren e Fredriks (2001), o gráfico *Worm-plot*, assim como o *Q-Q plot*, utiliza a comparação entre quantis observados e quantis teóricos para avaliar,

de forma visual, a adequação da suposição de normalidade dos resíduos. Esse gráfico, no entanto, facilita a identificação de regiões específicas em que o modelo viola tal suposição. Espera-se que os pontos se distribuam de maneira aproximadamente linear em torno da linha de referência, linha vermelha, sem apresentar curvaturas, permanecendo dentro dos limites de confiança.

2.8 Seleção de Modelos

O processo de seleção de modelos corresponde à etapa em que se busca identificar, entre as alternativas disponíveis, o modelo que melhor descreve a relação entre a variável resposta e os preditores. Em outras palavras, procura-se selecionar o modelo mais parcimonioso, isto é, aquele que apresenta menor complexidade sem perda significativa de capacidade explicativa.

Com esse objetivo, o Critério de Informação de Akaike (AIC), proposto por Akaike (1974), visa selecionar modelos bem ajustados com o menor número possível de parâmetros, minimizando a divergência de Kullback–Leibler (KL), introduzida por Kullback e Leibler (1951). Essa divergência mede a distância entre o modelo estatístico ajustado e o modelo verdadeiro, de modo que valores menores indicam menor perda de informação. Assim, o AIC é definido por:

$$\text{AIC} = -2\, l(\hat{\boldsymbol{\beta}}; \mathbf{y}) + 2p,$$

em que $l(\hat{\boldsymbol{\beta}}; \mathbf{y})$ denota o valor da log-verossimilhança maximizada e p representa o número de parâmetros do modelo (Paula, 2013).

Um segundo critério amplamente utilizado é o Critério de Informação Bayesiano (BIC), também conhecido como critério de Schwarz, proposto por Schwarz (1978). Esse critério busca equilibrar a qualidade do ajuste, medida pela verossimilhança, e a parcimônia do modelo, penalizando modelos mais complexos. O BIC pode ser interpretado como uma aproximação assintótica da probabilidade a posteriori do modelo. Sua expressão é dada por

$$\text{BIC} = -2\, l(\hat{\boldsymbol{\beta}}; \mathbf{y}) + p\, \log(n),$$

em que, novamente, $l(\hat{\boldsymbol{\beta}}; \mathbf{y})$ representa o valor da log-verossimilhança maximizada, p é o número de parâmetros do modelo e n denota o tamanho da amostra.

Os critérios AIC e BIC destacam-se por não necessitarem de testes estatísticos em seu processo de seleção de modelos. Além desses, existem outros procedimentos de seleção, como por exemplo o critério C_p de Mallows (Mallows, 1973), o método *stepwise* (Hocking, 1976), bem como os métdos de *backward* e *forward*.

2.9 Teste de Vuong

O teste de *Vuong* é utilizado para a comparação de modelos não aninhados, isto é, modelos que não são casos particulares um do outro (Vuong, 1989). Viabilizando a identificação da especificação mais compatível com o processo de geração dos dados em estudo. A base do teste consiste na comparação das log-verossimilhanças individuais dos modelos para cada observação. Com foco em cada observação individual i, calcula-se

$$m_i = \log f_1(y_i \mid x_i) - \log f_2(y_i \mid x_i).$$

Essa expressão representa a diferença entre as log-verossimilhanças do modelo 1 em relação ao modelo 2. Dessa forma, resultados positivos de m_i indicam um melhor ajuste para o modelo 1 em comparação ao modelo 2, enquanto valores negativos evidenciam maior consistência para o modelo 2 em relação ao modelo 1.

A estatística do teste é definida por

$$\mathrm{V} = \frac{\sqrt{n}\,\bar{m}}{s_m},$$

em que n é o tamanho da amostra, $\bar{m}$ é a média das diferenças de log-verossimilhança e s_m é o desvio padrão amostral dos m_i. Sob a hipótese nula $\mathrm{H}_0 : \mathbb{E}(m_i) = 0$ de que ambos os modelos são igualmente próximos do verdadeira distribuição geradora dos dados, e hipótese alternativa de que $\mathrm{H}_1 : \mathbb{E}(m_i) > 0$, indicando melhor ajuste do modelo 1, e $\mathrm{H}_1 : \mathbb{E}(m_i) < 0$, indicando adequação mais robusta do modelo 2.

Sob a hipótese nula, a estatística V converge em distribuição para uma Normal padrão. Em um teste unilateral, o valor-p é calculado como valor-$p = 1 - \Phi(\mathrm{V})$ quando a hipótese alternativa é $\mathrm{H}_1 : \mathbb{E}(m_i) > 0$, com valor-$p = \Phi(\mathrm{V})$ quando $\mathrm{H}_1 : \mathbb{E}(m_i) < 0$, sepresenta $\Phi(\mathrm{V})$ expressa a função de distribuição acumulada da Normal padrão. Para testes bilaterais o valor-p é denotado por valor-$p = 2\{1 - \Phi(|\mathrm{V}|)\}$ (Vuong, 1989).

3 MODELOS PARA DADOS DE CONTAGEM

3.1 Introdução

Dados de contagem correspondem a variáveis que assumem apenas valores inteiros não negativos. Esse tipo de variável é comum em diversas aplicações nas quais se deseja modelar o número de ocorrências de um evento em função de covariáveis explicativas. Exemplos frequentes incluem o número de prisões registradas para um indivíduo ao longo de um ano, a quantidade de atendimentos de emergência relacionados ao uso de drogas em uma semana, o número de cigarros consumidos diariamente ou ainda o total de patentes registradas por uma empresa em um determinado período (Wooldridge, 2001).

O termo "dados de contagem" refere-se a um conjunto de observações associadas a eventos ou itens que podem ser enumerados. No contexto estatístico, esse tipo de dado corresponde a valores inteiros não negativos, iniciando em zero e se estendendo até algum valor superior, que pode não ser previamente determinados. Embora, em teoria, as contagens possam assumir valores de zero até o infinito, na prática elas são limitadas por um valor máximo observado no conjunto de dados analisado, o qual representa o maior número de ocorrências registrado no estudo (Hilbe, 2014).

Tradicionalmente, a distribuição Poisson é utilizada para modelar dados de contagem pois possui propriedades que se adequam bem a essa classe de dados, como por exemplo independência de eventos, igualdade entre média e variância no conjunto de dados e sua interpretação intuitiva. Porém, quando a amostra apresenta variabilidade maior que a esperada pelo modelo, indicando superdispersão, as pressuposições fundamentais da distribuição Poisson são violadas.

A incidência e a intensidade da superdispersão variam de acordo com o contexto de aplicação. Diversos fatores podem contribuir para a ocorrência, incluindo características do processo de coleta de dados, heterogeneidade não observada e excesso de zeros. Como consequência, a não acomodação da superdispersão pode conduzir a erros-padrão elevados, estimativas imprecisas e valores de desvio inconsistêntes, o que tende a favorecer a escolha de modelos mais complexos (Cordeiro; Demétrio, 2008).

A regressão binomial negativo é amplamente empregada para modelar dados de contagem originalmente ajustáveis por Poisson, mas que apresentam superdispersão. Atualmente, ela é reconhecida como a abordagem padrão para lidar com situações em que

dados de contagem apresentam variabilidade superior à esperada, especialmente quando a origem da superdispersão não é conhecida (Hilbe, 2007).

Contudo, alguns conjuntos de dados apresentam uma quantidade de zeros maior do que a esperada sob uma modelagem Poisson ou Binomial Negativo. Com o objetivo de acomodar esse excesso de zeros, foi proposto por Lambert (1992) os modelos inflacionados de zeros, que permitem modelar explicitamente essa característica. As principais abordagens dessa classe são os modelos ZIP e ZINB, os quais possibilitam a distinção e modelagem de entre zeros estruturais e zeros amostrais. Um exemplo específico é a contagem de lesões de doenças em plantas: uma planta pode não apresentar lesões por ser resistente à doença ou simplesmente porque nenhum esporo do patógeno a atingiu (Ridout *et al.*, 1998).

3.2 Distribuição Poisson

A distribuição Poisson é um distribuição de probabilidade discreta utilizada para modelar o número de observações de um evento em um determinado intervalo de tempo, área ou volume, desenvolvida por Poisson (1837). Suponha uma conjunto de variáveis aleatórias Y_i, independentes entre si, utilizadas para modelagem de dados de contagem com suporte nos número naturais incluindo o zero, com taxa média de ocorrência constante, representada por μ_i e que apresente equidispersão do seu conjunto de dados, isto é média e variância equivalentes, com função de probabilidade dada por

$$\mathbb{P}(Y = y) = \frac{e^{-\mu_i}\mu_i^y}{y!} \tag{3.1}$$

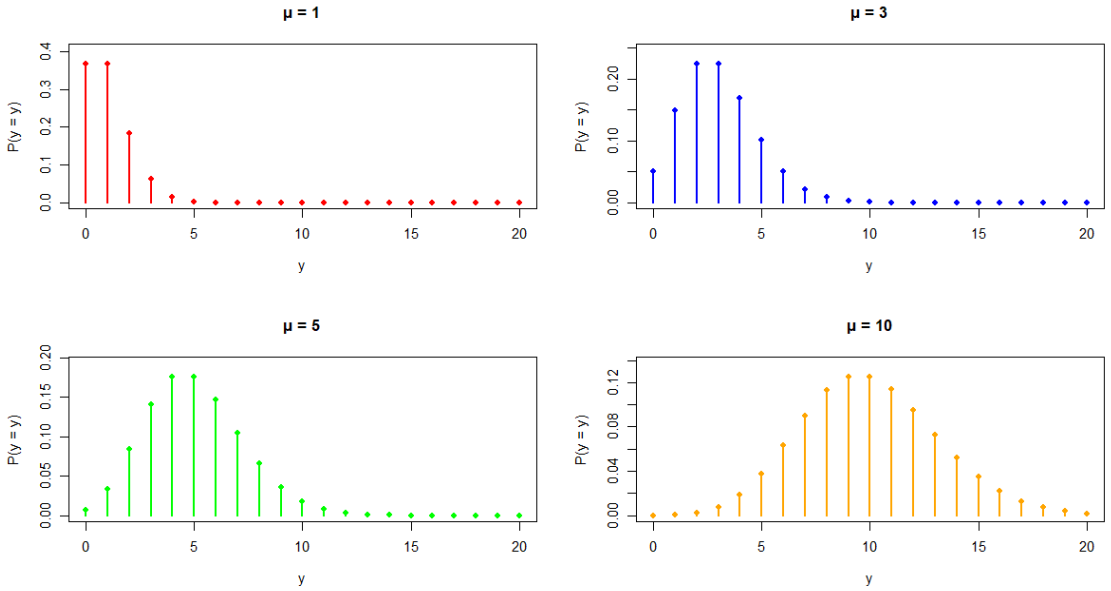
em que $y = 0, 1, \ldots$ e $\mu > 0$.

A distribuição representa uma legítima distribuição de probabilidade, pois

$$\sum_{y=0}^{\infty} \mathbb{P}(Y = y) = \sum_{y=0}^{\infty} \frac{e^{-\mu}\mu^y}{y!} = e^{-\mu} \sum_{y=0}^{\infty} \frac{\mu^y}{y!} = e^{-\mu}e^{\mu} = e^0 = 1.$$

Na Figura 1 apresenta-se a função de probabilidade da distribuição Poisson para diferentes valores do parâmetro μ.

Figura 1 – Função de probabilidade da distribuição Poisson para diferentes valores do parâmetro μ.



Se $Y \sim \text{Poisson}(\mu)$ com parâmetro μ então a média $\mathbb{E}(Y) = \mu$ e variância $\text{Var}(Y) = \mu$.

Demonstração: Pela definição de esperação

$$\mathbb{E}(Y) = \sum_{y=0}^{\infty} y \frac{e^{-\mu_i}\mu_i^y}{y!} = \sum_{y=1}^{\infty} \frac{e^{-\mu_i}\mu_i^y}{(y-1)!},$$

fazendo $x = y - 1$, verificamos que

$$\sum_{x=0}^{\infty} \frac{e^{-\mu}\mu_i^{x+1}}{x!} = \mu_i \sum_{x=0}^{\infty} \frac{e^{-\mu}\mu^x}{x!} = \mu.$$

O segundo momento é obtido da seguinte forma

$$\mathbb{E}(Y^2) = \sum_{y=0}^{\infty} y^2 \frac{e^{-\mu_i}\mu_i^y}{y!} = \sum_{y=1}^{\infty} y \frac{e^{-\mu_i}\mu_i^y}{(y-1)!},$$

fazendo novamente $x = y - 1$, verificamos que

$$\sum_{x=0}^{\infty} (x+1) \frac{e^{-\mu_i}\mu_i^{x+1}}{x!} = \mu_i \sum_{x=0}^{\infty} x_i \frac{e^{-\mu_i}\mu_i^x}{x!} + \mu \sum_{x=0}^{\infty} \frac{e^{-\mu_i}\mu_i^x}{x!} = \mu_i^2 + \mu_i.$$

Dessa forma, a variância de Y é expressa por

$$\text{Var}(Y) = \mathbb{E}(Y^2) - \mathbb{E}^2(Y) = \mu_i^2 + \mu_i - \mu_i^2 = \mu_i.$$

3.3 Modelo Poisson

Sejam $Y_1, Y_2, \ldots, Y_n$ variáveis aleatórias independentes condicionadas a um vetor de covariáveis $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\top$, com $i = 1, 2, \ldots, n$, também conhecido como vetor de variáveis explicativas. Considere que a distribuição condicional é $Y_i \mid \mathbf{x}_i \sim \text{Poisson}(\mu_i)$, cuja função de probabilidade é dada por

$$f(y_i \mid \mathbf{x}_i) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}, \tag{3.2}$$

em que $\mu_i > 0$ e $y_i = 0, 1, 2, \ldots$. Essa distribuição pertence à família exponencial linear, pois pode ser escrita como

$$\begin{aligned} f(y_i \mid \mathbf{x}_i) &= \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!} \\ &= \exp\{-\mu_i + y_i \log(\mu_i) - \log(y_i!)\} \\ &= \exp\{y_i \log(\mu_i) - \mu_i - \log(y_i!)\}, \end{aligned}$$

sendo o parâmetro de dispersão $\phi = 1$, a função de ligação canônica $\theta_i = \log(\mu_i)$, com $b(\theta_i) = e^{\theta_i}$ e $c(y_i; \phi) = -\log(y_i!)$. Assim, a média condicional, conforme (2.2), é expressa por

$$\mathbb{E}(Y_i \mid \mathbf{x}_i) = \frac{d}{d\theta_i}\left(e^{\theta_i}\right) = e^{\theta_i} = e^{\log(\mu_i)} = \mu_i, \tag{3.3}$$

e a variância condicional, conforme (2.3), é dada por

$$\text{Var}(Y_i \mid \mathbf{x}_i) = \frac{d}{d\theta_i}\left(e^{\theta_i}\right) = e^{\theta_i} = e^{\log(\mu_i)} = \mu_i. \tag{3.4}$$

Portanto, tem-se $V(\mu_i) = \mu_i$. A relação com o preditor linear é dada por

$$\eta_i = \log(\mu_i) = \mathbf{x}_i^\top \boldsymbol{\beta}, \tag{3.5}$$

em que $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^\top$ é um vetor de parâmetros de dimensão $p \times 1$ (Paula, 2013).

Assim, o modelo com média condicional dada em (3.3) e variância condicional dada em (3.4) apresenta a propriedade de heterocedasticidade, uma vez que a variância condicional da variável resposta não é constante ao longo das observações, característica compartilhada por todos os modelos de contagem (Ramalho, 1996).

3.3.1 Resíduos para o modelo Poisson

Nessa capitulo, apresentaremos os resíduos discutidos na seção (2.7) adaptados ao modelo Poisson, além de expor a expressão da *Deviance* correspondente.

Resíduos de Pearson

Tomando como base a expressão geral em (2.5) e o estimador da média e da variância da Poisson apresentada em (3.3) e (3.4), respectivamente, temos os resíduos de Pearson para o modelo Poisson, denotado por

$$r_{P_i} = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}}.$$

Resíduos de Pearson Padronizado

Cabe ressaltar que o cálculo dos resíduos de Pearson padronizados requer a obtenção da matriz de projeção **H** descrita em (2.7), composta pela matriz **X** de especificação do modelo e a matriz **W** conhecida como matriz de pesos. Já para o modelo Poisson com função de ligação canônica expressão por $\eta_i = \log(\mu_i)$, o i-ésimo elemento da matriz **W** fica denotado por

$$w_i = \left(\frac{d\mu_i}{d\eta_i}\right)^2 V_i^{-1} = \left(\frac{d}{d\eta_i} e^{\eta_i}\right)^2 \frac{1}{\mu_i} = \frac{(e^{\eta_i})^2}{\mu_i} = \frac{\mu_i^2}{\mu_i} = \mu_i,$$

assim, a expressão para o resíduos de Pearson padronizados conforme (2.6) fica denotado pela expressão

$$r_{P_i^*} = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i(1 - \hat{h}_{ii})}}.$$

Resíduos de *Deviance*

Os resíduos de *Deviance* para o modelo Poisson apresentam uma caracteristica de separação em duas partes quando $y_i = 0$ e $y_i > 0$, pois surge uma expressão $\log(0)$ durante seu desenvolvimento, cujo valor é indefinido. Dessa forma a separação em partes se torna nescessária. Expressamos a função log-verossimilhança para o modelo Poisson a partir do função de probabilidade apresentada em (3.2) por

$$l(\boldsymbol{\mu_i}|y_i) = \sum_{i=1}^{n} \left[y_i \log(\mu_i) - \mu_i - \log(y_i!) \right],$$

assim, utilizando a definição apresentada em (2.4), temos que

$$\begin{aligned} D^*(\mathbf{y}; \hat{\boldsymbol{\mu}}) &= 2 \sum_{i=1}^{n} \left[y_i \log(y_i) - y_i - \log(y_i!) - (y_i \log(\hat{\mu}_i) - \hat{\mu}_i - \log(y_i!)) \right] \\ &= 2 \sum_{i=1}^{n} \left[y_i \log(y_i) - y_i - y_i \log(\hat{\mu}_i) + \hat{\mu}_i \right] \\ &= 2 \sum_{i=1}^{n} \left[y_i \log\left(\frac{y_i}{\hat{\mu}_i} \right) - (y_i - \hat{\mu}_i) \right]. \end{aligned}$$

Essa é a expressão para o caso com $y_i > 0$. Já para o caso em que $y_i = 0$ a função de probabilidade fica denotada por

$$f(y_i = 0|\mathbf{x}_i) = \frac{e^{-\mu_i} \mu_i^0}{0!} = e^{-\mu_i},$$

portanto, a log-verossimilhança fica definia da seguinte forma

$$l(\boldsymbol{\mu_i}|y_i) = \log(e^{-\mu_i}).$$

Assim, a expressão dos resíduos de *Deviance* é dada por

$$
\begin{aligned}
D^*(\mathbf{y};\hat{\boldsymbol{\mu}}) &= 2\sum_{i=1}^{n}\left[\log(e^{-y_i}) - \log(e^{-\hat{\mu}_i})\right] \\
&= 2\sum_{i=1}^{n}\left[\log(e^{0}) - \log(e^{-\hat{\mu}_i})\right] \\
&= 2\sum_{i=1}^{n}\mu_i.
\end{aligned}
$$

Dessa forma, conforme (2.8) a forma conjunta dos resíduos de *Deviance* para o modelo Poisson ficam especificada por

$$
r_{D_i} = \begin{cases} \text{sinal}(y_i - \hat{\mu}_i)\sqrt{2\hat{\mu}_i}, & \text{se } y_i = 0 \\ \text{sinal}(y_i - \hat{\mu}_i)\sqrt{2\left[y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i - \hat{\mu}_i)\right]}, & \text{se } y_i > 0 \end{cases}
$$

a função sinal$(\cdot)$ indica o sinal da diferença entre o valor observado e o valor ajustado.

Resíduos de *Deviance* Padronizado

Apresentando os resíduos de *Deviance* em sua forma padronizada, conforme a expressão em (2.9), obtém-se a seguinte formulação

$$
r_{D_i} = \begin{cases} \dfrac{\text{sinal}(y_i - \hat{\mu}_i)\sqrt{2\hat{\mu}_i}}{\sqrt{1-\hat{h}_{ii}}}, & \text{se } y_i = 0, \\ \dfrac{\text{sinal}(y_i - \hat{\mu}_i)\sqrt{2\left[y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i - \hat{\mu}_i)\right]}}{\sqrt{1-\hat{h}_{ii}}}, & \text{se } y_i > 0. \end{cases}
$$

de forma que $\hat{h}_{ii}$ consiste no i-ésimo elemento da matriz $\hat{\mathbf{H}}$, denominada matriz de projeção.

3.4 Distribuição Binomial Negativo

A distribuição de probabilidade Binomial Negativo, desenvolvida por Montmort (1713), é uma distribuição discreta que descreve o número de tentativas necessárias até que ocorra um número pré-determinado de sucessos em experimentos independentes de Bernoulli.

Em cada tentativa, há uma probabilidade p de sucesso e $1-p$ de fracasso. Assim, essa distribuição pode ser interpretada de duas formas, o número total de tentativas necessárias para alcançar o r-ésimo sucesso ou, alternativamente, o número de fracassos observados antes de se atingir o r-ésimo sucesso, sendo esta última interpretação conhecida como distribuição de Pascal.

Seja Y uma variável aleatória que segue uma distribuição Binomial Negativo com número de fracassos até o r-ésimo sucesso, com função deprobablidade definida por

$$\mathbb{P}(Y=y)=\binom{y+r-1}{r-1}p^r(1-p)^y \tag{3.6}$$

em que $y=0,1,2,\ldots$, a distribuição representa uma legítima função de probabilidade, pois

$$\sum_{y=0}^{\infty}\mathbb{P}(Y=y)=\sum_{y=0}^{\infty}\binom{y+r-1}{r-1}p^r(1-p)^y=p^r\sum_{y=0}^{\infty}\binom{y+r-1}{r-1}(1-p)^y$$

utilizando a identidade da série binomial negativo

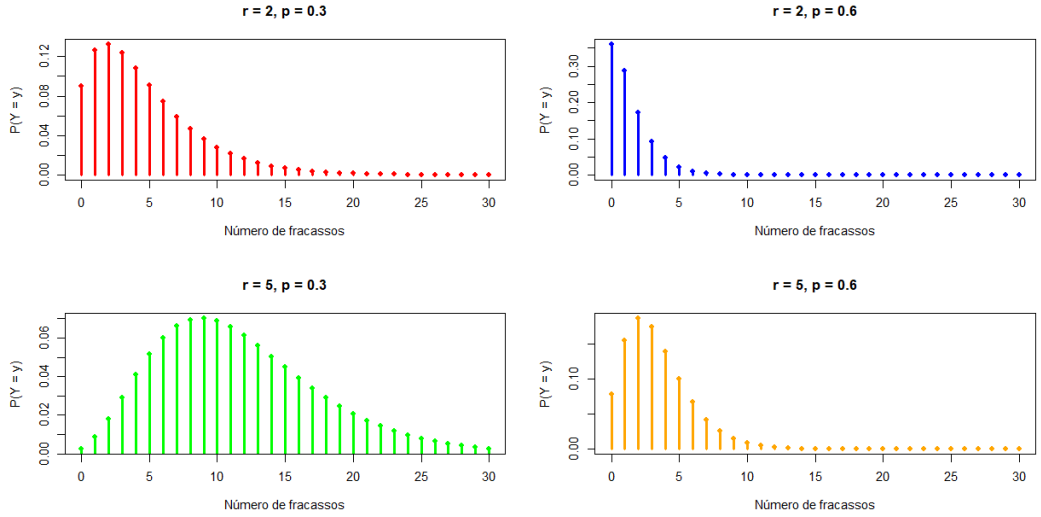
$$\sum_{k=0}^{\infty}\binom{k+r-1}{r-1}x^k=\frac{1}{(1-x)^r}\quad\text{para } |x|<1,$$

temos que

$$\sum_{y=0}^{\infty}\mathbb{P}(Y=y)=p^r\sum_{y=0}^{\infty}\binom{y+r-1}{r-1}(1-p)^y=p^r\frac{1}{(1-(1-p))^r}=\frac{p^r}{p^r}=1.$$

Na Figura 2, apresenta-se a função de probabilidade da distribuição Binomial Negativo, considerando o número de fracassos até o r-ésimo sucesso, para diferentes valores dos parâmetros.

Figura 2 – Função de probabilidade da distribuição Binomial negativo considerando a quantidade de fracassos.



Se $Y \sim \textit{Binomial Negativo}(r,\ p)$ então

$$\mathbb{E}(Y) = \frac{r(1-p)}{p} \qquad e \qquad \mathrm{Var}(Y) = \frac{r(1-p)}{p^2}.$$

Demonstração: Pela definição

$$\mathbb{E}(Y) = \sum_{y=0}^{\infty} y \binom{y+r-1}{r-1} p^r (1-p)^y,$$

consideramos a série geradora

$$S(x) = \sum_{y=0}^{\infty} \binom{y+r-1}{r-1} x^y = (1-x)^{-r} \qquad |x| < 1.$$

Note que, para $S(x) = (1-x)^{-r}$, temos

$$p^r \sum_{y=0}^{\infty} y \binom{y+r-1}{r-1} x^y = p^r x S^{(1)}(x) = x \cdot p^r r (1-x)^{-r-1},$$

substituindo $x = 1-p$, temos

$$p^r \left[(1-p)\, r(1-(1-p))^{-r-1}\right] = p^r (1-p)\, r p^{-r-1} = \frac{r(1-p)}{p}.$$

Para obter o segundo momento primeiramente calculamos $\mathbb{E}(Y(Y-1))$. Usando a segunda derivada de $S(x)$,

$$\sum_{y=0}^{\infty} y(y-1) \binom{y+r-1}{r-1} x^y = x^2 S''(x) = x^2 r(r+1)(1-x)^{-r-2},$$

logo, para $x = 1 - p$,

$$\mathbb{E}(Y(Y-1)) = p^r\,(1-p)^2 r(r+1)p^{-r-2} = \frac{r(r+1)(1-p)^2}{p^2},$$

em seguida

$$\mathbb{E}(Y^2) = \mathbb{E}(Y(Y-1)) + \mathbb{E}(Y) = \frac{r(r+1)(1-p)^2}{p^2} + \frac{r(1-p)}{p},$$

e colocando em denominador comum p^2,

$$\mathbb{E}(Y^2) = \frac{r(r+1)(1-p)^2 + r(1-p)p}{p^2} = \frac{r(1-p)\big((r+1)(1-p)+p\big)}{p^2}.$$

Por fim,

$$\mathrm{Var}(Y) = \mathbb{E}(Y^2) - \mathbb{E}^2(Y) = \frac{r(1-p)\big((r+1)(1-p)+p\big)}{p^2} - \frac{r^2(1-p)^2}{p^2},$$

portanto

$$\mathrm{Var}(Y) = \frac{r(1-p)}{p^2}.$$

Seja Y uma variável aleatória que segue uma distribuição Binomial Negativo considerando o número de tentativas até o r-ésimo sucesso p, sua função de probablidade, definida por

$$\mathbb{P}(Y = y) = \binom{y-1}{r-1} p^r (1-p)^{y-r} \tag{3.7}$$

em que $y = r, r+1, r+2, \ldots$.

A distribuição representa uma função de probabilidade, pois

$$\sum_{y=r}^{\infty} \mathbb{P}(Y = y) = \sum_{y=r}^{\infty} \binom{y-1}{r-1} p^r\,(1-p)^{y-r},$$

e fazendo $m = y - r$ a soma converte-se em

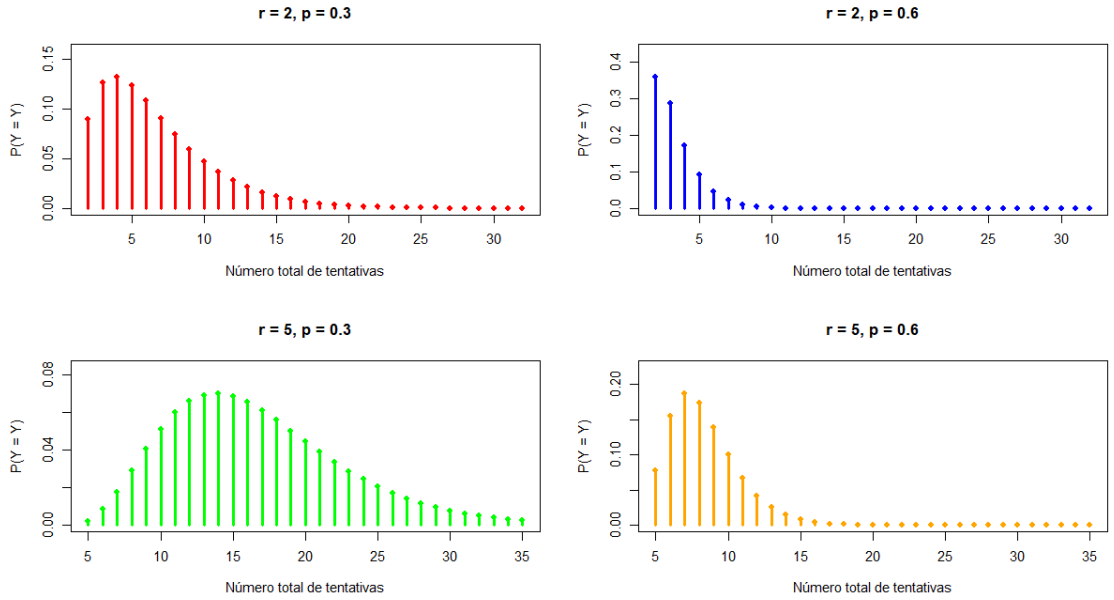
$$p^r \sum_{m=0}^{\infty} \binom{r+m-1}{m} (1-p)^m,$$

utilizando a identidade da série binomial negativo

$$p^r \sum_{m=0}^{\infty} \binom{r+m-1}{m} (1-p)^m = p^r \frac{1}{p^r} = 1.$$

Na Figura 3, é apresentada-se a função de probabilidade da distribuição Binomial Negativo, considerando o número de tentativas até o r-ésimo sucesso, para diferentes valores dos parâmetros.

Figura 3 – Função de probabilidade da distribuição Binomial negativo considerando a quantidade de tentativas.



Se $Y \sim Binomial\ Negativo(r,\ p)$, então

$$\mathbb{E}(Y) = \frac{r}{p}\ e\ \mathrm{Var}(Y) = \frac{r(1-p)}{p^2}$$

Demonstração: Pela definição

$$\mathbb{E}(Y) = \sum_{y=r}^{\infty} y \binom{y_i - 1}{r-1}\ p^r (1-p)^{y-r},$$

substituindo k = y - r, temos

$$\sum_{k=0}^{\infty} (k+r) \binom{k+r-1}{r-1} p^r (1-p)^k,$$

$$= rp^r \sum_{k=0}^{\infty} \binom{k+r-1}{r-1} (1-p)^k + p^r \sum_{k=0}^{\infty} k \binom{k+r-1}{r-1} (1-p)^k,$$

utilizando a identidade de séries binomial negativo

$$\sum_{k=0}^{\infty} \binom{k+r-1}{r-1} (1-p)^k = \frac{1}{p^r},$$

e

$$\sum_{k=0}^{\infty} k\binom{k+r-1}{r-1}(1-p)^k = \frac{r(1-p)}{p^{r+1}}.$$

retornando a expressão para $\mathbb{E}[Y]$

$$\mathbb{E}(Y) = \frac{rp^r}{p^r} + \frac{p^r r(1-p)}{p^{r+1}} = r + \frac{r(1-p)}{p} = r\left(1 + \frac{1-p}{p}\right) = \frac{r}{p}.$$

O segundo momento é obtido com o segundo momento fatorial da seguinte forma

$$\mathbb{E}(Y(Y-1)) = p^r \sum_{y=r}^{\infty} y(y-1)\binom{y-1}{r-1}(1-p)^{y-r},$$

alternando-se a variável $k = y - r$, temos

$$\mathbb{E}(Y(Y-1)) = p^r \sum_{k=0}^{\infty}(k+r)(k+r-1)\binom{k+r-1}{r-1}(1-p)^k,$$

$$= p^r\left(\sum_{k=0}^{\infty} k(k-1)\binom{k+r-1}{r-1}(1-p)^k\right.$$

$$+2r\sum_{k=0}^{\infty} k\binom{k+r-1}{r-1}(1-p)^k$$

$$\left.+r(r-1)\sum_{k=0}^{\infty}\binom{k+r-1}{r-1}(1-p)^k\right),$$

e utilizando a identidade de séries binomial negativo

$$\sum_{k=0}^{\infty}\binom{k+r-1}{r-1}(1-p)^k = \frac{1}{p^r},$$

$$\sum_{k=0}^{\infty} k\binom{k+r-1}{r-1}(1-p)^k = \frac{r(1-p)}{p^{r+1}},$$

$$\sum_{k=0}^{\infty} k(k-1)\binom{k+r-1}{r-1}(1-p)^k = \frac{r(r+1)(1-p)^2}{p^{r+2}},$$

e substituindo-se na expressão inicial, tem-se:

$$\mathbb{E}(Y(Y-1)) = p^r \left(\frac{r(r+1)(1-p)^2}{p^{r+2}} + 2r \cdot \frac{r(1-p)}{p^{r+1}} + r(r-1) \cdot \frac{1}{p^r} \right)$$

$$= \frac{r(r+1)(1-p)^2}{p^2} + \frac{2r^2(1-p)}{p} + r(r-1).$$

Para o segundo momento $\mathbb{E}[Y^2] = \mathbb{E}(Y(Y-1)) + \mathbb{E}(Y)$

$$\mathbb{E}(Y^2) = \frac{r(r+1)(1-p)^2}{p^2} + \frac{2r^2(1-p)}{p} + r(r-1) + \frac{r}{p},$$

$$\mathrm{Var}(Y) = \mathbb{E}(Y^2) - \mathbb{E}^2(Y) = \left(\frac{r(r+1)(1-p)^2}{p^2} + \frac{2r^2(1-p)}{p} + r(r-1) + \frac{r}{p} \right) - \frac{r^2}{p^2}$$

$$= \frac{r(r+1)(1-p)^2 + 2r^2p(1-p) + r(r-1)p^2 + rp - r^2}{p^2},$$

De forma que

$$\mathrm{Var}(Y) = \frac{r(1-p)}{p^2}.$$

A distribuição Binomial Negativo também pode ser obtida como resultado de uma mistura Poisson–Gama. Nesse contexto, assume-se que a variável aleatória condicional $Y \mid \lambda$ segue uma distribuição de Poisson com parâmetro λ, enquanto o parâmetro λ segue uma distribuição Gama com parâmetros (α, β). Essa abordagem foi desenvolvida por Greenwood e Yule (1920) como consequência dos estudos de modelos aplicados a surtos de doenças e ao registro de acidentes industriais.

Considere, portanto, λ uma variável aleatória com distribuição Gama, de parâmetros α e β, cuja função densidade de probabilidade é dada por

$$f(\lambda; \alpha, \beta) = \frac{\left(\frac{\beta}{\alpha}\right)^{\beta}}{\Gamma(\beta)} \lambda^{\beta-1} \exp\left\{ -\frac{\beta\lambda}{\alpha} \right\}, \quad \lambda > 0, \tag{3.8}$$

com $\alpha, \beta > 0$. Assumindo que $Y \mid \lambda \sim \text{Poisson}(\lambda)$ e que $\lambda \sim \text{Gama}(\alpha, \beta)$, a função de probabilidade marginal de Y pode ser obtida a partir da distribuição condicional por meio da integração em relação a λ, isto é,

$$
\begin{aligned}
f(y) &= \int_0^\infty \frac{e^{-\lambda}\lambda^y}{y!} \cdot \frac{\left(\frac{\beta}{\alpha}\right)^\beta}{\Gamma(\beta)} \lambda^{\beta-1} \exp\left\{-\frac{\beta\lambda}{\alpha}\right\} d\lambda \\
&= \frac{1}{y!\Gamma(\beta)} \left(\frac{\beta}{\alpha}\right)^\beta \int_0^\infty \lambda^y \lambda^{\beta-1} e^{-\lambda} \exp\left\{-\frac{\beta\lambda}{\alpha}\right\} d\lambda \\
&= \frac{1}{y!\Gamma(\beta)} \left(\frac{\beta}{\alpha}\right)^\beta \int_0^\infty \lambda^{y+\beta-1} \exp\left\{-\lambda\left(\frac{\beta}{\alpha}+1\right)\right\} d\lambda
\end{aligned}
$$

e considerando que $\Gamma(n) = \int_0^\infty x^{n-1}e^{-x}dx$ e fazendo a substituição $u = \lambda\left(\frac{\beta}{\alpha}+1\right)$, temos

$$
\begin{aligned}
f(y) &= \frac{1}{y!\Gamma(\beta)} \left(\frac{\beta}{\alpha}\right)^\beta \int_0^\infty \left(\frac{u}{\frac{\beta}{\alpha}+1}\right)^{y+\beta-1} e^{-u} \frac{du}{\left(\frac{\beta}{\alpha}+1\right)} \\
&= \frac{1}{y!\Gamma(\beta)\left(\frac{\beta}{\alpha}+1\right)^{y+\beta}} \left(\frac{\beta}{\alpha}\right)^\beta \int_0^\infty u^{y+\beta-1} e^{-u} du \\
&= \frac{\Gamma(y+\beta)}{y!\Gamma(\beta)\left(\frac{\beta}{\alpha}+1\right)^{y+\beta}} \left(\frac{\beta}{\alpha}\right)^\beta \\
&= \frac{\Gamma(y+\beta)}{\Gamma(y+1)\Gamma(\beta)} \left(\frac{\beta}{\alpha+\beta}\right)^\beta \left(\frac{\alpha}{\alpha+\beta}\right)^y
\end{aligned}
$$

logo,

$$
f(y;\alpha,\beta) = \frac{\Gamma(y+\beta)}{\Gamma(y+1)\Gamma(\beta)} \left(\frac{\beta}{\alpha+\beta}\right)^\beta \left(\frac{\alpha}{\alpha+\beta}\right)^y \tag{3.9}
$$

com $y = 0, 1, 2, \ldots$ e $\alpha, \beta > 0$, assim $Y \sim \textit{Binomial Negativo}(\alpha, \beta)$.

3.5 Modelo Binomial Negativo

Em dados de contagem, o modelo Binomial Negativo constitui uma alternativa ao modelo de Poisson quando o conjunto de dados apresenta superdispersão. Nessa modelagem, utiliza-se a parametrização da distribuição Binomial Negativo baseada no número de fracassos até o r-ésimo sucesso, com uma reparametrização em termos da média.

Para essa parametrização, tem-se

$$
\mathbb{E}(Y_i \mid \mathbf{x}_i) = \mu_i = \frac{r(1-p)}{p} \quad \Rightarrow \quad p = \frac{r}{r+\mu_i}.
$$

Sejam $Y_1, Y_2, \ldots, Y_n$ variáveis aleatórias independentes condicionadas a um vetor de covariáveis $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\top$, com $i = 1, 2, \ldots, n$, também conhecido como

vetor de variáveis explicativas. Assume-se que a distribuição condicional é dada por

$$Y_i \mid \mathbf{x}_i \sim \text{Binomial Negativo}(\mu_i, \phi),$$

cuja função de probabilidade é dada por

$$f(y_i|x_i) = \frac{\Gamma(y_i+\phi)}{\Gamma(y_i+1)\Gamma(\phi)} \left(\frac{\phi}{\mu_i+\phi}\right)^{\phi} \left(\frac{\mu_i}{\mu_i+\phi}\right)^{y_i}, \tag{3.10}$$

em que $y = 0, 1, \ldots$, ϕ é o parâmetro de dispersão e $\Gamma(\cdot)$ é a função Gama. Portanto, a Binomial Negativo pertencendo a família exponencial de distribuições, temos:

$$\begin{aligned}
f(y_i|x_i) &= \frac{\Gamma(y_i+\phi)}{\Gamma(y_i+1)\Gamma(\phi)} \left(\frac{\phi}{\mu_i+\phi}\right)^{\phi} \left(\frac{\mu_i}{\mu_i+\phi}\right)^{y_i} \\
&= \exp\left\{\log\left[\frac{\Gamma(y_i+\phi)}{\Gamma(y_i+1)\Gamma(\phi)} \left(\frac{\phi}{\mu_i+\phi}\right)^{\phi} \left(\frac{\mu_i}{\mu_i+\phi}\right)^{y_i}\right]\right\} \\
&= \exp\left\{\log\left[\frac{\Gamma(y_i+\phi)}{\Gamma(y_i+1)\Gamma(\phi)}\right] + \log\left[\left(\frac{\phi}{\mu_i+\phi}\right)^{\phi}\right] + \log\left[\left(\frac{\mu_i}{\mu_i+\phi}\right)^{y_i}\right]\right\} \\
&= \exp\left\{\log\left[\frac{\Gamma(y_i+\phi)}{\Gamma(y_i+1)\Gamma(\phi)}\right] + \phi\log\left(\frac{\phi}{\mu_i+\phi}\right) + y_i\log\left(\frac{\mu_i}{\mu_i+\phi}\right)\right\} \\
&= \exp\left\{y_i\log\left(\frac{\mu_i}{\mu_i+\phi}\right) + \phi\log\left(\frac{\phi}{\mu_i+\phi}\right) + \log\left[\frac{\Gamma(y_i+\phi)}{\Gamma(y_i+1)\Gamma(\phi)}\right]\right\},
\end{aligned}$$

a função de ligação canônica é representada por $\theta_i = \log(\frac{\mu_i}{\mu_i+\phi})$, $b(\theta_i) = -\phi\log(1-e^{\theta_i})$ e $c(y_i;\phi) = \frac{\Gamma(y_i+\phi)}{\Gamma(y_i+1)\Gamma(\phi)}$, dessa forma, a média condicional a partir de (2.2) é expressa por

$$\mathbb{E}(Y_i|x_i) = \frac{d}{d\theta_i}\left(-\phi\log(1-e^{\theta_i})\right) = \frac{\phi e^{\theta_i}}{1-e^{\theta_i}} = \mu_i, \tag{3.11}$$

e sua referente variância conforme (2.3) é expressa por

$$\text{Var}(Y_i|x_i) = \frac{d}{d\theta_i}\left(\frac{\phi e^{\theta_i}}{1-e^{\theta_i}}\right) = \frac{\phi e^{\theta_i}}{(1-e^{\theta_i})^2} = \mu_i + \frac{\mu_i^2}{\phi}, \tag{3.12}$$

evidenciando que $\text{Var}(Y_i \mid \mathbf{x}_i) > \mathbb{E}(Y_i \mid \mathbf{x}_i)$.

Além disso, quando $\phi \to \infty$, a distribuição Binomial Negativo converge para a distribuição de Poisson com média μ_i (Rodrigues, 2012). Tem-se que

$$f(y_i \mid \mathbf{x}_i) = \frac{\Gamma(y_i+\phi)}{\Gamma(y_i+1)\Gamma(\phi)} \left(\frac{\phi}{\mu_i+\phi}\right)^{\phi} \left(\frac{\mu_i}{\mu_i+\phi}\right)^{y_i}.$$

Observa-se que

$$\frac{\mu_i}{\mu_i + \phi} = \frac{1}{1 + \frac{\phi}{\mu_i}}, \quad \Gamma(y_i + 1) = y_i!,$$

de modo que a função de probabilidade pode ser reescrita como

$$f(y_i \mid \mathbf{x}_i) = \frac{\mu_i^{y_i}}{y_i!} \frac{\Gamma(y_i + \phi)}{\Gamma(\phi)\, \phi^{y_i}} \left(1 + \frac{\mu_i}{\phi}\right)^{-\phi}.$$

Fazendo $\phi \to \infty$, o segundo termo converge para 1 e o terceiro termo converge para $e^{-\mu_i}$, resultando em

$$f(y_i \mid \mathbf{x}_i) = \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!},$$

que corresponde à função de probabilidade da distribuição de Poisson com média μ_i.

Dessa forma, a distribuição Binomial Negativo apresenta-se como uma alternativa consistente ao modelo de Poisson, uma vez que se aproxima dessa distribuição quando ϕ tende ao infinito e, para valores finitos de ϕ, permite modelar situações em que a variância é superior à média.

3.5.1 Resíduos para o modelo Binomial Negativo

Nessa seção, apresentaremos os resíduos discutidos na seção (2.7) adaptados ao modelo Binomial Negativo, além de expor a expressão da *Deviance* correspondente.

Resíduos de Pearson

Aplicando a expressão geral (2.5) à estimativa da média e da variância da distribuição Binomial Negativo, descritas em (3.11) e (3.12), respectivamente, obtêm-se os resíduos de Pearson do modelo Binomial Negativo, definido por

$$r_{P_i} = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i + \frac{\hat{\mu}_i^2}{\phi}}}.$$

Resíduos de Pearson Padronizados

É importante observar que o cálculo dos resíduos de Pearson padronizados exige a determinação da matriz de projeção $\mathbf{H}$ apresentada em (2.7), formada pela matriz

de especificação do modelo $\mathbf{X}$ e pela matriz de pesos $\mathbf{W}$. No caso do modelo Binomial Negativo, com função de ligação canônica representada por $\eta_i = \log\left(\frac{\mu_i}{\mu_i - \phi}\right)$, o i-ésimo elemento da matriz $\mathbf{W}$ é expresso por

$$\begin{aligned}
w_i &= \left(\frac{d\mu_i}{d\eta_i}\right)^2 V_i^{-1} = \left[\frac{d}{d\eta_i}\left(\frac{e^{\eta_i}}{1-e^{\eta_i}}\right)\right]^2 \frac{1}{\mu_i + \frac{\mu_i^2}{\phi}} \\
&= \left[\frac{e^{\eta_i}(1-e^{\eta_i}) - (e^{\eta_i}(-e^{\eta_i}))}{(1-e^{\eta_i})^2}\right]^2 \frac{1}{\mu_i + \frac{\mu_i^2}{\phi}} \\
&= \left[\frac{e^{\log\left(\frac{\mu_i}{\mu_i+\phi}\right)}}{\left(1 - e^{\log\left(\frac{\mu_i}{\mu_i+\phi}\right)}\right)^2}\right]^2 \frac{\phi}{\mu_i(\mu_i+\phi)} \\
&= \left[\frac{\frac{\mu_i}{\mu_i+\phi}}{\left(1 - \frac{\mu_i}{\mu_i+\phi}\right)^2}\right]^2 \frac{\phi}{\mu_i(\mu_i+\phi)} \\
&= \frac{(\mu_i(\mu_i+\phi))^2}{\phi^2} \frac{\phi}{\mu_i(\mu_i+\phi)} = \frac{\mu_i(\mu_i+\phi)}{\phi}.
\end{aligned}$$

Assim, a expressão para o resíduos de Pearson padronizados conforme (2.7) é expresso por

$$r_{P_i^*} = \frac{y_i - \hat{\mu}_i}{\sqrt{\frac{\hat{\mu}_i(\hat{\mu}_i+\phi)}{\phi}(1-\hat{h}_{ii})}}.$$

Resíduos de *Deviance*

Os resíduos de *Deviance* para o modelo Binomial Negativo apresentam características semelhantes às do modelo Poisson, especialmente quanto à necessidade de separar a *Deviance* em dois casos $y_i = 0$ e $y_i > 0$. Essa separação é necessária, pois em determinadas etapas do desenvolvimento analítico, surgem termos do tipo $\log(0)$, que são indefinidos, tornando essencial a distinção entre os dois casos para garantir consistência matemática.

A função log-verossimilhança para o modelo Binomial Negativo pode ser obtida a partir da função de probabilidade apresentada em (3.10) e é dada por

$$l(\boldsymbol{\mu_i}|y_i) = \sum_{i=1}^{n}\left[y_i\log(\mu_i) + \phi\log(\phi) - (y_i+\phi)\log(\mu_i+\phi) + \log\left(\frac{\Gamma(y_i+\phi)}{\Gamma(y_i+1)\Gamma(\phi)}\right)\right]$$

e, utilizando a definição apresentada em (2.4), temos que

$$\begin{aligned}
D^*(\mathbf{y};\hat{\boldsymbol{\mu}}) &= 2\sum_{i=1}^{n}\left[y_i\log(y_i)+\phi\log(\phi)-(y_i+\phi)\log(y_i+\phi)+\log\left(\frac{\Gamma(y_i+\phi)}{\Gamma(y_i+1)\Gamma(\phi)}\right)\right.\\
&\quad\left.-\left(y_i\log(\hat{\mu}_i)+\phi\log(\phi)-(y_i+\phi)\log(\hat{\mu}_i+\phi)+\log\left(\frac{\Gamma(y_i+\phi)}{\Gamma(y_i+1)\Gamma(\phi)}\right)\right)\right]\\
&= 2\sum_{i=1}^{n}\left[y_i\log(y_i)-y_i\log(\hat{\mu}_i)-(y_i+\phi)\log(y_i+\phi)+(y_i+\phi)\log(\hat{\mu}_i+\phi)\right]\\
&= 2\sum_{i=1}^{n}\left[y_i\log\left(\frac{y_i}{\hat{\mu}_i}\right)+\phi\log\left(\frac{\hat{\mu}_i+\phi}{y_i+\phi}\right)+y_i\log\left(\frac{\hat{\mu}_i+\phi}{y_i+\phi}\right)\right]\\
&= 2\sum_{i=1}^{n}\left[\phi\log\left(\frac{\hat{\mu}_i+\phi}{y_i+\phi}\right)+y_i\log\left(\frac{y_i(\hat{\mu}_i+\phi)}{\hat{\mu}_i(y_i+\phi)}\right)\right],
\end{aligned}$$

que é a expressão para o caso com $y_i > 0$. Para o caso com $y_i = 0$ a função de probabilidade fica denotada por

$$f(y_i=0|\mathbf{x}_i)=\frac{\Gamma(0+\phi)}{\Gamma(0+1)\Gamma(\phi)}\left(\frac{\phi}{\mu_i+\phi}\right)^{\phi}\left(\frac{\mu_i}{\mu_i+r}\right)^{0}=\left(\frac{\phi}{\mu_i+\phi}\right)^{\phi},$$

com log-verossimilhança definida da seguinte forma

$$l(\boldsymbol{\mu_i}|y_i)=\left[\phi\log(\phi)-\phi\log(\mu_i+\phi)\right].$$

Assim, os resíduos de *Deviance* ficam definidos da seguinte forma

$$\begin{aligned}
D^*(\mathbf{y};\hat{\boldsymbol{\mu}}) &= 2\sum_{i=1}^{n}\Big[\phi\log\phi-\phi\log(y_i+\phi)\\
&\quad-\big(\phi\log\phi-\phi\log(\hat{\mu}_i+\phi)\big)\Big]\\
&= 2\sum_{i=1}^{n}\big[-\phi\log(\phi)+\phi\log(\hat{\mu}_i+\phi)\big]\\
&= 2\sum_{i=1}^{n}\phi\log\left(\frac{\hat{\mu}_i+\phi}{\phi}\right).
\end{aligned}$$

Dessa forma, conforme (2.7) a expressão para o resíduo *Deviance* para o modelo Binomial Negativo é especificado por

$$r_{D_i} = \begin{cases} \text{sinal}(y_i - \hat{\mu}_i)\sqrt{2\left[\phi \log\left(\frac{\hat{\mu}_i+\phi}{\phi}\right)\right]}, & \text{se } y_i = 0 \\ \text{sinal}(y_i - \hat{\mu}_i)\sqrt{2\left[\phi \log\left(\frac{\hat{\mu}_i+\phi}{y_i+\phi}\right) + y_i \log\left(\frac{y_i(\hat{\mu}_i+\phi)}{\hat{\mu}_i(y_i+\phi)}\right)\right]}, & \text{se } y_i > 0 \end{cases}$$

Resíduos de *Deviance* Padronizado

Os resíduos de *Deviance* na sua forma padronizada, de acordo com (3.2), são definidos por

$$r_{D_i} = \begin{cases} \dfrac{\text{sinal}(y_i - \hat{\mu}_i)\sqrt{2\left[\phi \log\left(\frac{\hat{\mu}_i+\phi}{\phi}\right)\right]}}{\sqrt{(1-\hat{h}_{ii})}}, & \text{se } y_i = 0, \\ \dfrac{\text{sinal}(y_i - \hat{\mu}_i)\sqrt{2\left[\phi \log\left(\frac{\hat{\mu}_i+\phi}{y_i+\phi}\right) + y_i \log\left(\frac{y_i(\hat{\mu}_i+\phi)}{\hat{\mu}_i(y_i+\phi)}\right)\right]}}{\sqrt{(1-\hat{h}_{ii})}}, & \text{se } y_i > 0. \end{cases}$$

de forma que $\hat{h}_{ii}$ é o i-ésimo elementa da matriz $\mathbf{H}$, denominada matriz de projeção.

3.6 Modelos de Contagem Inflacionados de Zeros

Os modelos de contagem inflacionados de zeros foram desenvolvidos por Lambert (1992) com o objetivo de modelar dados de contagem que apresentam excesso de zeros. Esse excesso pode ocorrer tanto em situações em que os zeros são gerados pela própria distribuição de contagem quanto na presença de zeros estruturais.

Como exemplo, considere o estudo do número de dias em que um indivíduo consome determinado produto em um período fixo. Indivíduos que não consomem o produto por razões como intolerância ou alergia produzem zeros estruturais, os quais não podem ser adequadamente modelados por distribuições de contagem tradicionais (Paula, 2013).

Seja Y_i uma variável aleatória com função de probabilidade dada por

$$\mathbb{P}\{Y_i = y_i\} = \begin{cases} \pi_i + (1-\pi_i)f(0,\theta_i), & \text{se } y = 0, \\ (1-\pi_i)f(y_i;\theta_i), & \text{se } y = 1, 2, \ldots \end{cases} \tag{3.13}$$

em que π_i denota a probabilidade de pertencer a um conjunto de zeros estruturais, assumindo valores no intervalo $0 < \pi_i < 1$ e $f(y_i;\theta_i)$ é a função de probabilidade da distribuição de contagem, por exemplo a Poisson, Binomial Negativo, pois

$$\begin{aligned}\sum_{y_i=0}^{\infty}\mathbb{P}\{Y_i=y_i\} &= \mathbb{P}\{Y_i=0\}+\sum_{y_i=1}^{\infty}\mathbb{P}\{Y_i=y_i\}\\ &= \pi_i+(1-\pi_i)f(0;\theta_i)+\sum_{y=1}^{\infty}(1-\pi_i)f(y_i;\theta_i)\\ &= \pi_i+(1-\pi)\left[f(0;\theta_i)+\sum_{y=1}^{\infty}f(y_i;\theta_i)\right]\\ &= \pi_i+(1-\pi_i)\left[f(0;\theta_i)+(1-f(0;\theta_i))\right]\\ &= \pi_i+(1-\pi_i)\cdot 1=1.\end{aligned}$$

Neste trabalho, consideraremos $Y\sim ZIP(\lambda)$ para o modelo Poisson inflacionado de zeros e $Y\sim ZINB(\mu,r)$ para o modelo Binomial Negativo inflacionado de zeros.

O primeiro e segundo momento da variável aleatória Y são expressos por

$$\begin{aligned}\mathbb{E}(Y_i) &= \sum_{y=1}^{\infty}y(1-\pi_i)f(y_i;\theta_i)\\ &= (1-\pi_i)\sum_{y=1}^{\infty}y_i f(y_i;\theta_i)\\ &= (1-\pi_i)\mathbb{E}(Y_i^*)\end{aligned}\tag{3.14}$$

e

$$\begin{aligned}\mathbb{E}(Y_i^2) &= \sum_{y=1}^{\infty}y_i^2(1-\pi_i)f(y_i;\theta_i)\\ &= (1-\pi_i)\sum_{y=1}^{\infty}y_i^2 f(y_i;\theta_i)\\ &= (1-\pi_i)\mathbb{E}[(Y_i^*)^2]\end{aligned}\tag{3.15}$$

em que Y_i^* representa a variável associada à distribuição de contagem base. E sua respectiva variância é representada por

A variância da variável aleatória Y_i é definida por

$$\mathrm{Var}(Y_i)=\mathbb{E}(Y_i^2)-\left[\mathbb{E}(Y_i)\right]^2.$$

Substituindo as expressões obtidas para o primeiro e o segundo momentos, tem-se

$$\begin{aligned}\mathrm{Var}(Y_i) &= (1-\pi_i)\mathbb{E}\big[(Y_i^*)^2\big] - \big[(1-\pi_i)\mathbb{E}(Y_i^*)\big]^2 \\ &= (1-\pi_i)\mathbb{E}\big[(Y_i^*)^2\big] - (1-\pi_i)^2\big[\mathbb{E}(Y_i^*)\big]^2,\end{aligned}$$

utilizando a identidade

$$\mathbb{E}\big[(Y_i^*)^2\big] = \mathrm{Var}(Y_i^*) + \big[\mathbb{E}(Y_i^*)\big]^2,$$

obtém-se

$$\begin{aligned}\mathrm{Var}(Y_i) &= (1-\pi_i)\left\{\mathrm{Var}(Y_i^*) + \big[\mathbb{E}(Y_i^*)\big]^2\right\} - (1-\pi_i)^2\big[\mathbb{E}(Y_i^*)\big]^2 \\ &= (1-\pi_i)\mathrm{Var}(Y_i^*) + \pi_i(1-\pi_i)\big[\mathbb{E}(Y_i^*)\big]^2,\end{aligned}$$

portanto,

$$\mathrm{Var}(Y_i) = (1-\pi_i)\mathrm{Var}(Y_i^*) + \pi_i(1-\pi_i)\big[\mathbb{E}(Y_i^*)\big]^2. \tag{3.16}$$

3.7 Modelo Poisson Inflacionado de Zeros (ZIP)

Em estudos com dados de contagem em que a frequência de zeros observada é maior do que a esperada pelo modelo Poisson tradicional, isto é, quando há a presença de zeros estruturais não provenientes de um processo aleatório inerente ao experimento e há evidência de superdispersão no conjunto de dados, é apropriado utilizar a modelagem ZIP. Considera-se que a variável resposta assume o valor zero com probabilidade π_i e, com probabilidade $(1-\pi_i)$, segue uma distribuição Poisson com média μ_i, conforme proposto por Lambert (1992). Assim, a função de probablidade é dada por:

$$\mathbb{P}\{Y_i = y_i\} = \begin{cases} \pi_i + (1-\pi_i)e^{-\mu_i}, & \text{se } y_i = 0, \\ (1-\pi_i)\frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}, & \text{se } y_i = 1, 2, \ldots. \end{cases} \tag{3.17}$$

Dessa forma, podemos obter a média e variância de (3.17) a partir de (3.14) e (3.16), respectivamente

$$\begin{aligned}\mathbb{E}(Y_i) &= (1-\pi_i)\mathbb{E}(Y_i^*) \\ &= (1-\pi_i)\mu_i\end{aligned} \tag{3.18}$$

e

$$\begin{aligned} \text{Var}(Y_i) &= (1-\pi_i)\text{Var}(Y_i^*) + \pi_i(1-\pi_i)\big[\mathbb{E}(Y_i^*)\big]^2 \\ &= (1-\pi_i)\mu_i + \pi_i(1-\pi_i)\mu_i^2 \end{aligned} \tag{3.19}$$

Utilizamos a teoria dos MLGs para estimar os parâmetros π_i e μ_i de forma que para o ajuste do modelo ZIP utilizamos a função de ligação em duas partes. Primeiramente a parte de contagem e na segunda a parte inflacionada, respectivamente, expressas por

$$\log \mu_i = \mathbf{X}\boldsymbol{\beta} \quad \text{e} \quad \log\left(\frac{\pi_i}{1-\pi_i}\right) = \mathbf{G}\boldsymbol{\gamma}$$

sendo $\mathbf{X}$ e $\mathbf{G}$ as matrizes de covariáveis associadas ao modelo em questão, e $\boldsymbol{\beta}$ e $\boldsymbol{\gamma}$ são vetores de parâmetros que podem ou não coincidir (Ridout *et al.*, 1998).

3.7.1 Resíduos para o Modelo ZIP

Nessa subseção, apresentaremos os resíduos discutidos na seção (2.7) adaptados ao modelo ZIP, além de expor a expressão da *Deviance* correspondente.

Resíduos de Pearson

Tomando como base a expressão geral em (2.5) conjuntamente com a média e a variância do modelo ZIP, apresentados em (3.18) e (3.19), respectivamente, podemos concluir que os resíduos de Pearson para o modelo ZIP ficam denotados por

$$r_{P_i} = \frac{y_i - (1-\hat{\pi}_i)\hat{\mu}_i}{\sqrt{(1-\hat{\pi}_i)\hat{\mu}_i + \hat{\pi}_i(1-\hat{\pi}_i)\hat{\mu}_i^2}}.$$

Resíduos de *Deviance*

Os resíduos de *Deviance* para o modelo ZIP são constituídos a partir da separação das observações para zeros estruturais e zeros amostrais. Dessa forma, a expressão da componente da *Deviance* para o caso $y > 0$, conforme (2.4), fica denotada por

$$\begin{aligned} d_i^2 &= 2[(-y_i + y_i \log y_i - \log y_i!) \\ &\quad - (\log(1-\hat{\pi}_i) - \hat{\mu}_i + y_i \log \hat{\mu}_i - \log y_i!)] \\ &= 2\left[y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i - \hat{\mu}_i) - \log(1-\hat{\pi}_i)\right] \end{aligned}$$

e de forma similar, encontramos a expressão para o caso $y = 0$

$$d_i^2 = -2\log\left(\hat{\pi}_i + (1-\hat{\pi}_i)e^{-\hat{\mu}_i}\right).$$

Com esse resultados, obtemos os resíduos de *Deviance* para o modelo ZIP:

$$r_{D_i} = \begin{cases} \text{sinal}(y_i - \hat{\mu}_i)\sqrt{-2\log\left[\hat{\pi}_i + (1-\hat{\pi}_i)e^{-\hat{\mu}_i}\right]}, & \text{se } y_i = 0, \\ \text{sinal}(y_i - \hat{\mu}_i)\sqrt{2\left[y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i - \hat{\mu}_i) - \log(1-\hat{\pi}_i)\right]}, & \text{se } y_i > 0. \end{cases}$$

3.8 Modelo Binomial Negativo Inflacionado de Zeros (ZINB)

Em estudos com dados de contagem nos quais a frequência de zeros observada é maior do que é comportado pelo modelo Binomial Negativo, isto é, quando há presença de zeros estruturais não provenientes do processo aleatório do experimento e, simultaneamente, há superdispersão no conjunto de dados, é apropriado utilizar a modelagem ZINB. Nesse modelo, considera-se que a variável resposta assume o valor zero com probabilidade π_i, correspondente aos zeros estruturais, e, com probabilidade $(1-\pi_i)$, segue uma distribuição Binomial Negativo com média μ_i e parâmetro de dispersão ϕ (Yau *et al.*, 2003).

$$\mathbb{P}(Y_i = y_i) = \begin{cases} \pi_i + (1-\pi_i)\left(\frac{\phi}{\phi+\mu_i}\right)^{\phi}, & y_i = 0, \\ (1-\pi_i)\frac{\Gamma(y_i+\phi)}{\Gamma(y_i+1)\Gamma(\phi)}\left(\frac{\phi}{\phi+\mu_i}\right)^{\phi}\left(\frac{\mu_i}{\phi+\mu_i}\right)^{y_i}, & y_i = 1, 2, \ldots \end{cases} \tag{3.20}$$

Dessa forma, a média e a variância de (3.20) podem ser obtidas de forma semelhante ao modelo ZIP. Seja Y_i^* uma variável aleatória com distribuição Binomial Negativo. Assim, dados por

$$\begin{aligned} \mathbb{E}(Y_i) &= (1-\pi_i)\mathbb{E}(Y_i^*) \\ &= (1-\pi_i)\mu_i, \end{aligned} \tag{3.21}$$

e

$$\begin{aligned}\mathrm{Var}(Y_i) &= (1-\pi_i)\,\mathrm{Var}(Y_i^*) + \pi_i(1-\pi_i)\left[\mathbb{E}(Y_i^*)\right]^2 \\ &= (1-\pi_i)\left(\mu_i + \frac{\mu_i^2}{\phi}\right) + \pi_i(1-\pi_i)\mu_i^2 \\ &= (1-\pi_i)\mu_i + (1-\pi_i)\mu_i^2\left(\pi_i + \frac{1}{\phi}\right).\end{aligned} \tag{3.22}$$

Tem-se que a distribuição ZINB, no limite em que $\phi \to \infty$, aproxima-se da distribuição ZIP, e, quando $\pi_i \to 0$, converge para a distribuição Binomial Negativo. Além disso, quando simultaneamente $1/\phi \approx 0$ e $\pi_i \approx 0$, a distribuição ZINB converge para a distribuição de Poisson (Montoya, 2009).

Dessa forma, o modelo ZINB é formulado a partir de duas componentes, de maneira análoga ao modelo ZIP: a primeira corresponde ao processo de contagem e a segunda à componente inflacionada de zeros, sendo ambas especificadas por funções de ligação próprias, dadas por

$$\log(\mu_i) = \mathbf{X}\boldsymbol{\beta} \qquad \text{e} \qquad \log\left(\frac{\pi_i}{1-\pi_i}\right) = \mathbf{G}\boldsymbol{\gamma},$$

em que $\mathbf{X}$ e $\mathbf{G}$ são matrizes de covariáveis associadas ao modelo e $\boldsymbol{\beta}$ e $\boldsymbol{\gamma}$ são vetores de parâmetros que podem ou não coincidir (Ridout *et al.*, 1998).

3.8.1 Resíduos para o Modelo ZINB

Nesta seção, apresentam-se os resíduos discutidos na Seção (2.7), adaptados ao modelo ZINB, bem como a expressão da *Deviance* correspondente.

Resíduos de Pearson

Tomando como base a expressão geral em (2.5) conjuntamente com a média e a variância do modelo ZINB apresentados em (3.21) e (3.22) respectivamente, podemos concluir que os resíduos de Pearson para o modelo ZINB ficam denotados por

$$r_{P_i} = \frac{y_i - (1-\hat{\pi}_i)\hat{\mu}_i}{\sqrt{(1-\hat{\pi}_i)\hat{\mu}_i + (1-\hat{\pi}_i)\hat{\mu}_i^2\left(\hat{\pi}_i + \frac{1}{\phi}\right)}}.$$

Resíduos de *Deviance*

Os resíduos de *Deviance* para o modelo ZINB são constituídos a partir da separação das observações para zeros estruturais e zeros amostrais, dessa forma a expressão dos componentes da *Deviance* para o caso $y_i > 0$ conforme (2.4) fica denotada por

$$d_i^2 = \begin{cases} -2\log\left[\hat{\pi}_i + (1-\hat{\pi}_i)\left(\frac{\phi}{\phi+\hat{\mu}_i}\right)^{\phi}\right], & \text{se } y_i = 0 \\ -2\left[\log(1-\hat{\pi}_i) + \log\left[\frac{\Gamma(y_i+\phi)}{\Gamma(y_i+1)\Gamma(\phi)}\left(\frac{\phi}{\hat{\mu}_i+\phi}\right)^{\phi}\left(\frac{\hat{\mu}_i}{\hat{\mu}_i+\phi}\right)^{y_i}\right]\right], & \text{se } y_i > 0 \end{cases}$$

com esses resultados obtemos os resíduos de *Deviance* para o modelo ZINB

$$r_{D_i} = \begin{cases} \text{sinal}(y_i-\hat{\mu}_i)\sqrt{-2\log\left[\hat{\pi}_i + (1-\hat{\pi}_i)\left(\frac{\phi}{\phi+\hat{\mu}_i}\right)^{\phi}\right]}, & \text{se } y_i = 0 \\ \text{sinal}(y_i-\hat{\mu}_i)\sqrt{2\left[y_i\log\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i+\phi)\log\left(\frac{y_i+\phi}{\hat{\mu}_i+\phi}\right) - \log(1-\hat{\pi}_i)\right]}, & \text{se } y_i > 0 \end{cases}$$

4 APLICAÇÃO

No presente capítulo será apresentada uma aplicação com dados oriundos de Atkins e Gallop (2007), os quais, por sua vez, provêm de um estudo anterior conduzido por Christensen *et al.* (2004). A análise foi inteiramente desenvolvida no *software* `R` (R Core Team, 2025), uma linguagem de programação de código aberto voltado à computação estatística, na versão 4.5.2, com auxílio do `RStudio` (Posit Software, PBC, 2025).

Foram ajustados modelos pertencentes à classe dos MLGs, incluindo as abordagens de Poisson e Binomial Negativo. Para cada um desses modelos, foi desenvolvido um código em R para a obtenção dos resíduos de Pearson padronizados, *Deviance* padronizados e quantílicos. De modo análogo, para os modelos de mistura ZIP e ZINB, também foram implementados códigos específicos para o cálculo dos resíduos de Pearson, *Deviance* e quantílicos. Por fim, realizou-se uma análise diagnóstica dos modelos, baseada na inspeção gráfica dos resíduos, contando com gráficos de HNP e Quantil-Quantil com envelopes simulados. Utilizou-se a função `hnp`() da biblioteca `HNP` e para os gráficos de *Worm-plot* empregou-se a função `wp`() da biblioteca `gamlss`, em conjunto com os resíduos obtidos para cada modelo, bem como no desenvolvimento de um código para cálculo da distância de Cook em conjunto com uma análise gráfica, com o objetivo de identificar o modelo que apresentou o melhor ajuste.

4.1 Terapia Conjugal

O conjunto de dados trata-se de uma amostra clínica composta por casais que buscaram terapia conjugal, totalizando 268 indivíduos, o que corresponde a 134 casais. A variável de interesse, MSI caracteriza-se por ser uma variável de contagem em uma escala de 0 a 14 que mensura o número de passos já considerados ou tomados em direção à separação ou ao divórcio. Para esse estudo foram consideradas as variáveis preditoras: *Dyadic Adjustment Scale* (DAS) expressa em uma escala de satisfação conjugal desenvolvida por Spanier (1976), variando de 0 a 151 e tratada como contínua; *Affective Communication* (AFC) uma medida contínua para problemas de comunicação afetiva, diferentemente da DAS, pontuações menores indicam melhor comunicação; *Sexual Dissatisfaction* (SEX) avalia o nível de descontentamento em relação à interação sexual do casal, também encarada como contínua; *Gender* é uma variável binária referente ao gênero 0 para mulher e 1 para

homem, por último, *Infidelity*, uma variável binária que indica a ocorrência de infidelidade no relacionamento com 0 para ausência e 1 para presença.

Os dados indicam que apenas 14.5% dos entrevistados relataram infidelidade na relação, como esperado, temos uma separação equivalente entre gêneros de participantes do estudo, como está apresentado na Tabela 2.

Tabela 2 – Descrição das variáves *Infidelity* e *Gender*

Variável	Sim %	Não %	Total
Infidelity	14,50	85,50	100,00
Gender	50,00	50,00	100,00

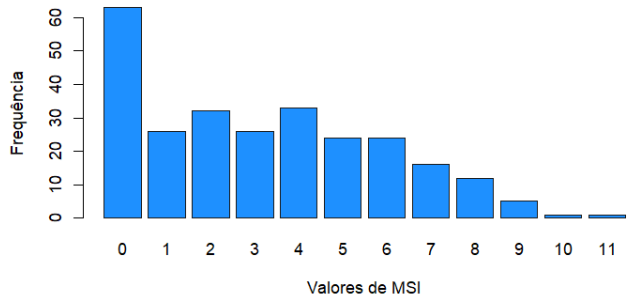
Avaliando as estatísticas descritivas para as variáveis quantitativas apresentadas na Tabela 3, podemos observar que a variável resposta MSI apresenta uma média relativamente baixa de 3.18 em contraste com uma variância substancialmente maior de 7,2056, evidenciando claramente uma superdispersão presente nos dados. Além disso fica evidente uma alta variância na variável DAS indicando uma heterogeneidade na satisfação conjugal dos indivíduos que participaram do estudo.

Tabela 3 – Descrição das variáveis Quantitativas

Variável	Min	1^o Q	Mediana	Média	3^o Q	Máx	Variância
MSI	0,00	1,00	3,00	3,18	5,00	11,00	7,2056
DAS	40,00	74,50	87,00	84,49	94,00	115,00	207,2586
AFC	45,00	59,00	63,00	63,35	69,00	76,00	46,9775

Na Figura 4 observamos a distribuição de frequência da variável explicativa MSI, indicando uma frequência de 63 zeros, equivalente a 24% da amostra, podendo ser o motivo da superdispersão apresentada nesse conjunto de dados. Dessa forma, é presumível que modelos convencionais não se adequem à base de dados.

Figura 4 – Distribuição da frequência da variável resposta MSI.



4.2 Modelo Poisson

Inicialmente desenvolve-se o modelo de regressão Poisson, por ser a primeira etapa numa estratégia de análise de dados de contagem, para o ajuste dos dados, conforme o modelo definido por Atkins e Gallop (2007). As variáveis contínuas DAS, AFC e SEX foram centradas em torno de suas respectivas médias, assim facilitando a interpretação do intercepto. Se as variáveis não fossem centradas, o intercepto representaria uma pessoa com pontuação zero, o que muitas vezes é um valor impossível e não representativo na amostra, também consideramos as variáveis binárias como categóricas. Dessa forma, utilizando a função logarítmica, temos

$$\begin{aligned}\log(\mu_i) = \beta_0 + \beta_1 DAS_i + \beta_2 AFC_i + \beta_3 SEX_i \\ + \beta_4 Infidelity_i + \beta_5 Gender_i + \beta_6 (DAS \times Infidelity)_i,\end{aligned}$$

nesse contexto, $i = 1, 2, \ldots, n$, μ_i representa o número médio de passos para o divórcio, bem como a interação DAS $\times$ *Infidelity* foi incluída no modelo para testar se o efeito da satisfação conjugal sobre os passos para o divórcio muda dependendo da ocorrência ou não de uma traição. Dessa forma, para o modelo Poisson no *software* `R`, utilizou-se a função `glm`() para modelagem dos dados. A Tabela 4 apresenta as estimativas e os erros-padrão do modelo Poisson.

Tabela 4 – Resumo da regressão do modelo Poisson para variável explicativa MSI

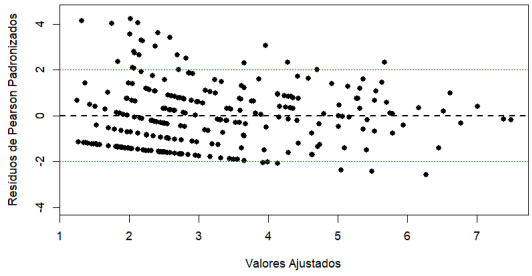
Parâmetros	$\hat{\beta}$	Erro-Padrão	Valor-P
Intercept	1,13	0,05	$< 0{,}001$
DAS	$-0{,}02$	0,00	$< 0{,}001$
AFC	0,02	0,01	$< 0{,}001$
SEX	0,01	0,01	0,127
Infidelity	0,53	0,09	$< 0{,}001$
Gender	$-0{,}21$	0,07	0,003
DAS $\times$ Infidelity	0,02	0,01	$< 0{,}001$

Dado o uso de uma função de ligação logarítmica, é necessário aplicar a exponencial aos coeficientes para interpretá-los na escala original da variável resposta. Consequentemente, devido à centralização dos preditores contínuos em torno da média, o intercepto no modelo representa o número esperado de passos em direção ao divórcio para uma pessoa que não teve um caso extraconjugal e que apresenta níveis médios de

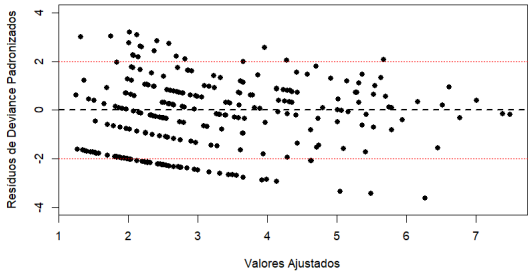
satisfação conjugal, insatisfação sexual e comunicação afetiva.

Dessa forma ao exponencializar o coeficiente do intercepto obtivemos que o modelo estima 3,1 passos em direção à separação e divórcio para esta combinação de preditores. Para casais sem infidelidade, cada aumento de uma unidade no DAS está associado a uma redução de aproximadamente 2% no número esperado de passos rumo ao divórcio. Cada aumento unitário em AFC está associado a um aumento de cerca de 2% no número esperado de passos em direção ao divórcio. Após o controle das demais variáveis, a variável SEX não apresenta associação estatisticamente significativa. Homens apresentam, em média, cerca de 19% menos passos em direção ao divórcio do que mulheres. Indivíduos que relataram infidelidade apresentam um aumento de aproximadamente 70% no número esperado de passos rumo ao divórcio, em comparação aos que não relataram infidelidade. Ainda, quando ocorre infidelidade, o nível geral de satisfação conjugal deixa de exercer um papel relevante sobre os comportamentos ligados à separação.

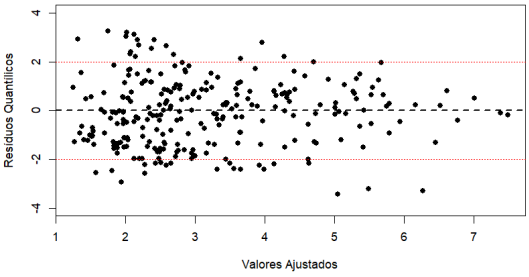
Figura 5 – Gráficos de resíduos vs valores ajustados - Modelo Poisson.



(a) Resíduos de Pearson padronizados



(b) Resíduos de *Deviance* padronizados

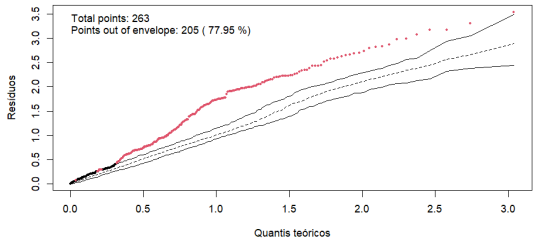


(c) Resíduos Quantílicos

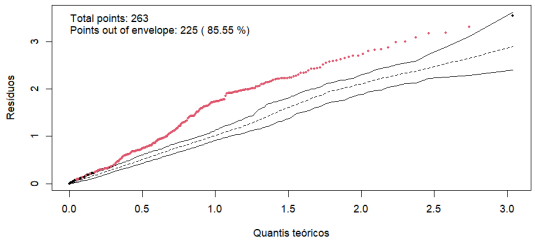
A análise dos gráficos de resíduos *vs* valores ajustados para o modelo de Poisson, apresentados na Figura 5, indicam que o modelo não apresenta padrões sistemáticos de não linearidade. Entretanto, verifica-se aumento da dispersão dos resíduos para os menores

valores ajustados, bem como a formação de estruturas em bandas, especialmente presentes no gráfico de resíduos de Pearson e *Deviance*, sendo uma característica de dados de contagem em que apresentam superdispersão. Já para os resíduos quantílicos Figura 5, embora apresentem comportamento mais próximo do esperado, ainda sugerem heterogeneidade da variância, isto é, a variância dos dados não está sendo corretamente modelada pela função de variância da Poisson. Essas evidências indicam uma violação das suposições de equidispersão intrínsecas do modelo Poisson, podendo ser causadas pelo excesso de zeros presentes na amostra.

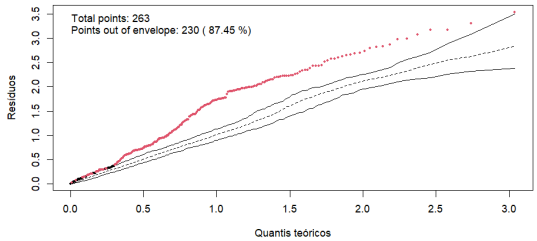
Figura 6 – Gráficos HNP com envelopes simulados - Modelo Poisson.



(a) Resíduos de Pearson padronizados.



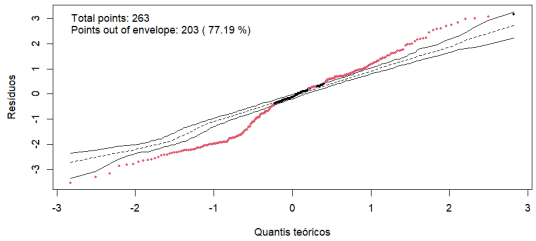
(b) Resíduos de *Deviance* padronizados.



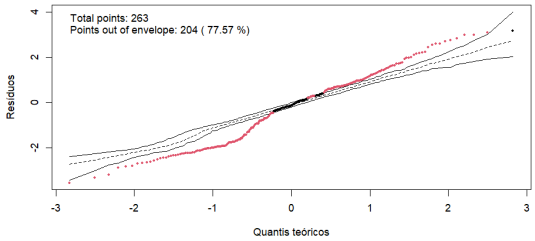
(c) Resíduos Quantílicos.

Seguindo com avaliação, os gráficos de HNP com envelopes simulados estão apresentados na Figura 6, evidenciando clara inadequação do modelo de Poisson aos dados analisados. Em todos os gráficos, Figura 6 percebe-se uma elevada quantidade de observações fora dos envelopes simulados, indicando um afastamento das curvas empíricas em relação ao padrão esperado da distribuição teórica. Assim, apontam um afastamento da variabilidade dos dados em relação ao esperado para o modelo Poisson, possivelmente causada pela presença de superdispersão e excesso de zeros presentes na amostra.

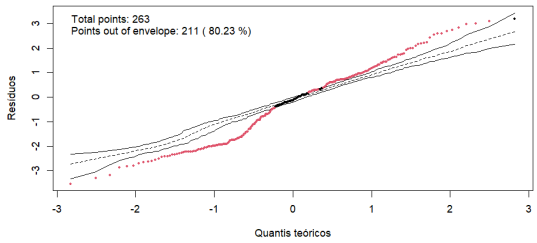
Figura 7 – Gráfico Quantis-Quantis com envelopes simulados - Modelo Poisson.



(a) Resíduos de Pearson padronizados.



(b) Resíduos de *Deviance* padronizados.

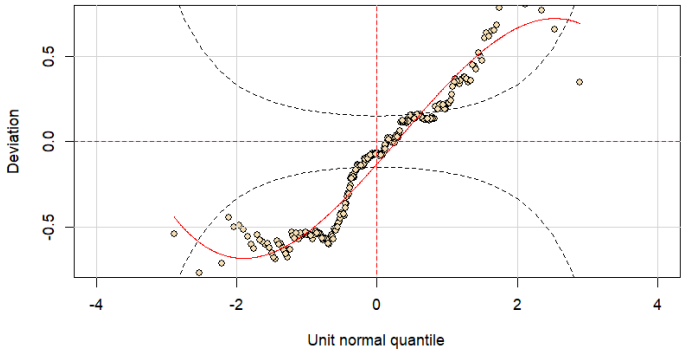


(c) Resíduos Quantílicos.

Prosseguindo com o procedimento analítico dos gráficos Quantil–Quantil com envelopes simulados dos resíduos de Pearson, *Deviance* e quantílicos, os mesmos são apresentados na Figura 7. Observa-se discrepâncias sistemáticas entre as distribuições empíricas dos resíduos e a distribuição teórica esperada sob o modelo de Poisson. Em conformidade, observa-se elevada quantidade de observações fora dos envelopes simulados, bem como desvios elevados nas caudas, indicando que a variabilidade dos dados não se adequa corretamente à suposição de equidispersão, reforçando a evidência de superdispersão e a necessidade de considerar modelos alternativos mais flexíveis.

Os gráficos *Worm-Plot* dos resíduos quantílicos, estão apresentados na Figura 8, deixando indícios de desvios regulares em relação ao comportamento esperado para o modelo de Poisson. Observam-se padrões nos resíduos, incluindo curvaturas em forma de "U" e "S", assimetria e a ocorrência de pontos fora das bandas de confiança, indicando problemas de escala associados à violação da suposição de equidispersão.

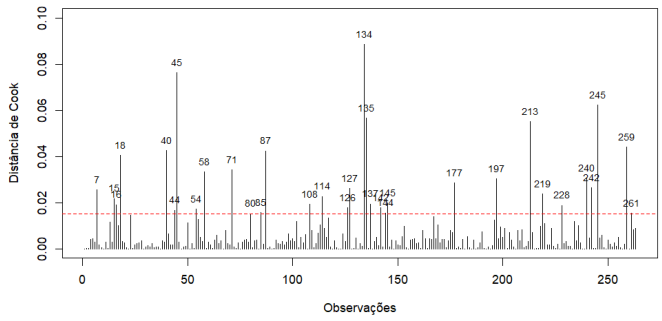
Figura 8 – Gráfico *Worm-Plot* para os resíduos Quantílicos - Modelo Poisson.



Porém, vale ressaltar que embora os resíduos de Pearson e *Deviance* possam ser usados na construção de gráficos de *Worm-plot*, sua interpretação deve ser realizada com cautela em modelos de contagem, uma vez que tais resíduos não apresentam, distribuição aproximadamente Normal padrão. Portanto, os resíduos quantílicos são mais apropriados para esse objetivo, pois são construídos de forma a seguir aproximadamente uma distribuição Normal padrão quando o modelo está corretamente especificado (Buuren; Fredriks, 2001).

O gráfico da Distância de Cook para o modelo de Poisson, apresentado na Figura 9, evidencia a presença de algumas observações potencialmente influentes, caracterizadas por valores superiores ao limiar adotado. Esses resultados indicam a não adequação às suposições estabelecidas para o modelo, já identificadas nos diagnósticos residuais realizados anteriormente. Tal comportamento sugere que os problemas de ajuste não decorre de observações influentes isoladas, mas estão associadas a limitações estruturais do modelo de Poisson, como a violação da suposição de equidispersão.

Figura 9 – Gráfico da Distância de Cook para o modelo Poisson.



Portanto, a análise conjunta dos diagnósticos realizados, indica de forma consistente que o modelo de Poisson não se ajusta adequadamente ao conjunto de dados analisado, assim as estimativas obtidas não são consideradas confiáveis. Diante dessas evidências, conclui-se que a distribuição de Poisson é restritiva para descrever a variabilidade nos dados, motivando a adoção de um modelo que incorpora um parâmetro adicional de dispersão, inicialmente tentaremos o modelo Binomial Negativo.

4.3 Modelo Binomial Negativo

Seguindo a mesma estratégia utilizada para a modelagem Poisson, adotou-se um procedimento análogo para o ajuste do modelo Binomial Negativo. O modelo foi ajustado considerando o mesmo conjunto de covariáveis e estrutura previamente adotados, utilizando a função `glm.nb()`, pertencente à biblioteca `MASS`. Isso foi feito de modo a permitir comparações diretas entre as abordagens, com objetivo de identificar se o parâmetro adicional de dispersão é capaz de acomodar a superdispersão presente no conjunto de dados.

$$\begin{aligned}\log(\mu_i) = \beta_0 + \beta_1 DAS_i + \beta_2 AFC_i + \beta_3 SEX_i \\ + \beta_4 Infidelity_i + \beta_5 Gender_i + \beta_6 (DAS_i \times Infidelity_i),\end{aligned}$$

Tabela 5 – Resumo da regressão do modelo Binomial Negativo.

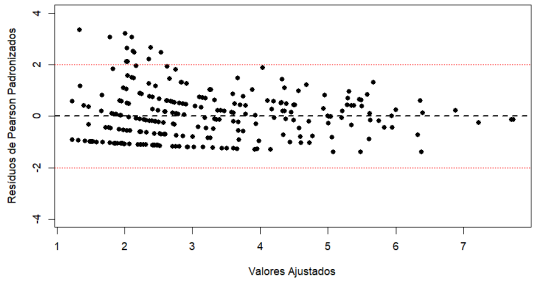
Parâmetros	$\hat{\beta}$	Erro-Padrão	Valor-P
Intercept	1,12	0,08	$< 0{,}001$
DAS	$-0{,}02$	0,00	$< 0{,}001$
AFC	0,02	0,01	0,031
SEX	0,01	0,01	0,310
Infidelity	0,51	0,15	$< 0{,}001$
Gender	$-0{,}20$	0,11	0,067
DAS $\times$ Infidelity	0,02	0,01	0,018

Na Tabela 5 estão apresentados as estimativas dos coeficientes do modelo em questão junto com os seus erros-padrão. Fica evidente uma relativa semelhança entre os coeficientes dos modelos, porém os erros-padrão são relativamente maiores para o modelo Binomial Negativo, isso se dá devido a acomodação de superdispersão proveniente do modelo em questão.

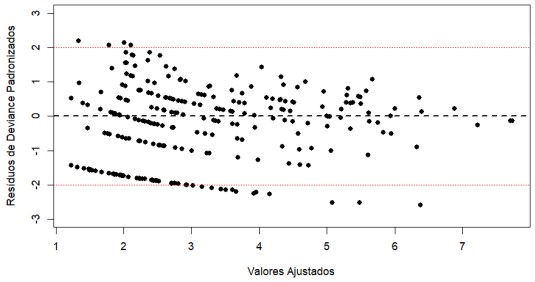
Em complemento, a análise dos gráficos de resíduos *vs* valores ajustados para

o modelo de Binomial Negativo, são apresentados na Figura 10, e apontam pouca melhora em comparação ao gráfico para o modelo Poisson, indicando que o a presença de excessos de zeros não estão sendo comportados tão bem pelo modelo Binomial Negativo.

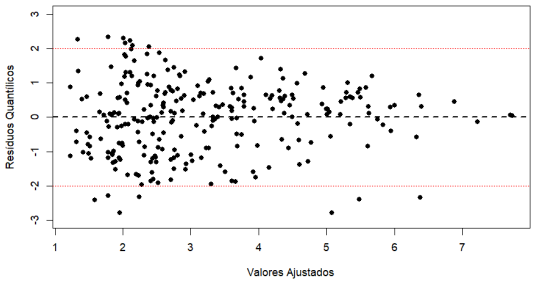
Figura 10 – Gráficos de resíduos *vs* valores ajustados - Modelo Binomial Negativo.



(a) Resíduos de Pearson padronizados.



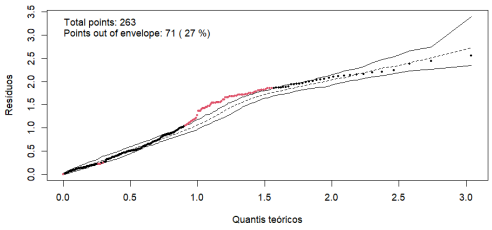
(b) Resíduos de *Deviance* padronizados.



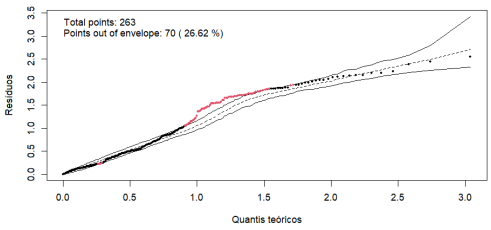
(c) Resíduos Quantílicos.

A análise dos gráficos HNP com envelopes simulados, apresentados na Figura 11, indicam uma melhora expressiva para o modelo Binomial Negativo em relação ao modelo Poisson, novamente evidenciando como o acréscimo do parâmetro de dispersão contribuiu com a melhora do modelo. Porém, a proporção de observações fora dos envelopes simulados ainda é significante, evidenciando que o modelo não comporta bem o excesso de zeros presente na amostra.

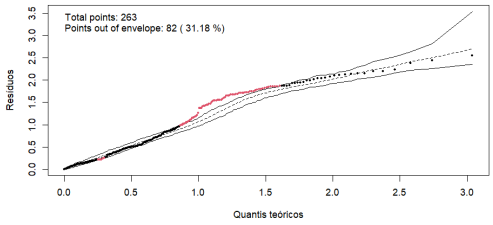
Figura 11 – Gráficos HNP com envelopes simulados - Modelo Binomial Negativo.



(a) Resíduos de Pearson padronizados.



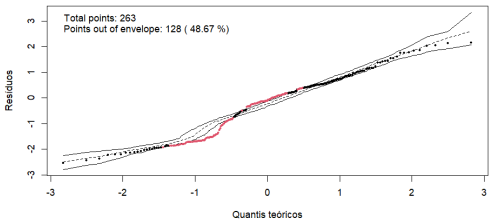
(b) Resíduos de *Deviance* padronizados.



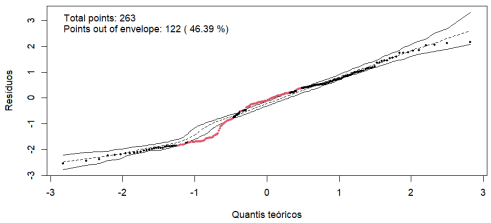
(c) Resíduos Quantílicos.

Em concordância os gráficos Quantil–Quantil com envelopes apresentados na Figura 12, indicam que o modelo apresenta comportamento residual considerávelmente mais próximo do esperado, apesar ainda serem observados determinados desvios localizados.

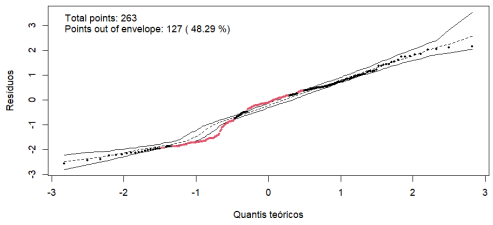
Figura 12 – Gráfico Quantis-Quantis com envelopes simulados - Modelo Binomial Negativo.



(a) Resíduos de Pearson padronizados.



(b) Resíduos de *Deviance* padronizados.

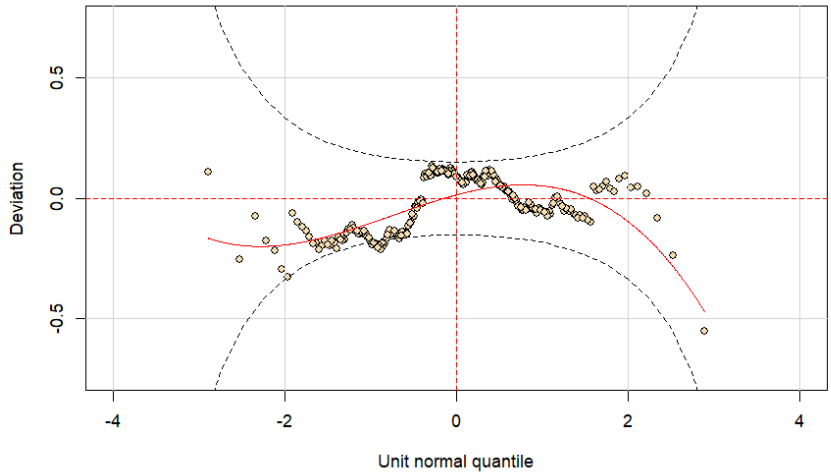


(c) Resíduos Quantílicos.

Em adição, o gráfico de *Worm-Plot* para os resíduos quantílicos, apresentado nas Figura 13, corrobora com as demais análises gráficas, indicando que o modelo Bino-

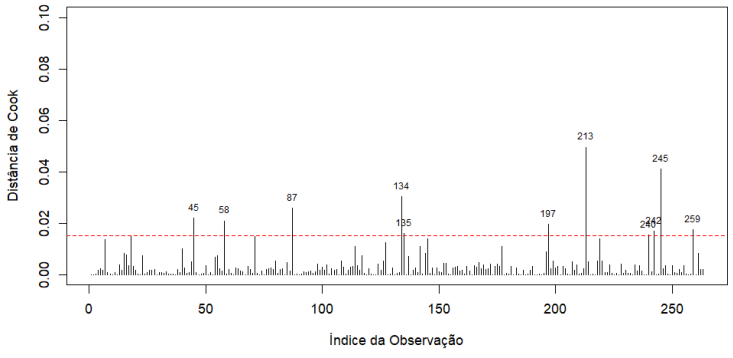
mial Negativo é capaz de acomodar, a superdispesão presente nos dados, pois mantem observações oscilando em torna da linha horizontal, com boa parte dentro das bandas de confiança, porém é evidente a não adequação ao excesso de zeros.

Figura 13 – Gráfico *Worm-Plot* para os resíduos Quantílicos - Modelo Binomial Negativo.



O gráfico da Distância de Cook para o modelo Binomial Negativo, apresentado na Figura 14, deixa evidente a existência de algumas observações potencialmente influentes, porém com quantidade inferior às observadas no modelo de Poisson. Assim, fica evidente uma melhora na estabilidade do ajuste, resultado das propriedades de adequação da superdispersão proveniente do modelo Binomial Negativo. Porém, em contraste o modelo Poisson apresentou maior sensibilidade às observações extremas, corroborando com a hipótese de inadequação do modelo Poisson estava ligada a não atendimento das suposições de equidispersão e não a presença de pontos influentes.

Figura 14 – Gráfico de Distância de Cook para o modelo Binomial Negativo.



Os resultados evidenciam uma melhora expressiva na qualidade do ajuste com a abordagem da Binomial Negativo em comparação ao modelo Poisson. Entretando, a presença de excesso de zeros nos dados ainda se torna um desafio para tal modelo, conforme evidenciado pela análise de diagnostico. Assim, torna-se necessário a utilização de modelos mais complexos como o ZIP e ZINB, especialmente projetados para comportar dados de contagem com excesso de zeros.

4.4 Modelo Poisson Inflacionado de Zeros

Diante da alta proporção de zeros presentes no conjunto de dados analisados, torna-se necessários a utilização de modelos mais flexível e complexos, capaz de acomodar essa característica. Nesse contexto, foi utilizado a abordagem ZIP, um modelo de mistura que combina uma componente responsável por contabilizar zeros estruturais e um processo de contagem usual, seguindo um distribuição Poisson.

A parte de contagem do modelo relaciona as variávies por meio da função de ligação logarítmica, enquanto isso a parte de zeros estruturais é modelada por meio de uma regressão logística, com função de ligação logit. Entretanto, a parte logit não modela o evento e sim a chance de pertencer ao grupo zero estrutural. Assim, consideramos a mesma abordagem de modelagem utilizada para a regressão Poisson quanto para a Binomial Negativo, utilizando a função `zeroinfl()` da biblioteca `pscl`. As estimativas dos coeficientes estão presentes na Tabela 6

Tabela 6 – Resumo da regressão do modelo ZIP.

Parâmetros	**Log (Contagem)**			**Logit (Zeros)**		
	$\hat{\beta}$	Erro-Padrão	**Valor-P**	$\hat{\beta}$	Erro-Padrão	**Valor-P**
Intercept	1,387	0,055	$< 0{,}001$	$-1{,}494$	0,273	$< 0{,}001$
DAS	$-0{,}010$	0,003	$< 0{,}001$	0,050	0,017	0,004
AFC	0,010	0,006	0,121	$-0{,}060$	0,029	0,035
SEX	0,003	0,004	0,423	$-0{,}019$	0,017	0,259
Gender	$-0{,}201$	0,075	0,007	$-0{,}053$	0,352	0,880
Infidelity	0,402	0,093	$< 0{,}001$	$-0{,}543$	0,669	0,417
DAS $\times$ Infidelity	0,008	0,005	0,121	$-0{,}080$	0,037	0,029

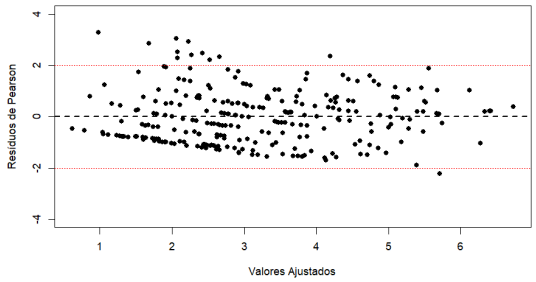
Dada a separação da amostra em parte de zeros estruturais e parte de contagem, também afeta a interpretação dos coeficientes gerados pela regressão. Para zeros estruturais um coeficiente positivo indica aumento na chance de pertencer a zeros estruturais e um

coeficiente negativo a chance de pertencer ao evento de contagem.

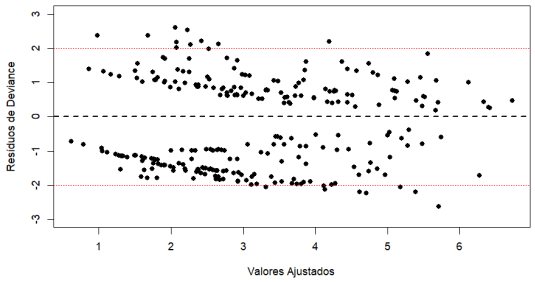
Nesse contexto, a estimativa para DAS indica que para cada aumento de uma unidade no DAS, as chances de não considerar o divórcio aumentam 5.1% . Problemas de comunicação afetiva AFC diminuem a chance de pertencer a zeros estruturais, aumentando os passos em direção ao divórcio conforme seu aumento. O abalo da infidelidade contribui para aproximar casais do divórcio, independente da qualidade do relacionamento.

Em relação à parte de contagem do modelo, maior satisfação conjugal está associada a uma redução de cerca de 1% na contagem de passos em direção ao divórcio entre aqueles que já os estão considerando. Homens que consideram o divórcio tendem a reportar cerca de 18% menos passos do que mulheres. Entre quem considera o divórcio, a infidelidade está associada a um aumento de quase 50% na contagem de passos em direção ao divórcio.

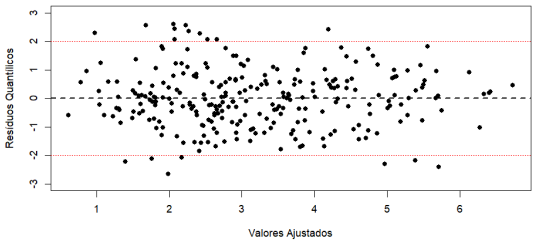
Figura 15 – Gráficos de resíduos *vs* valores ajustados - Modelo ZIP.



(a) Resíduos de Pearson



(b) Resíduos de *Deviance*.



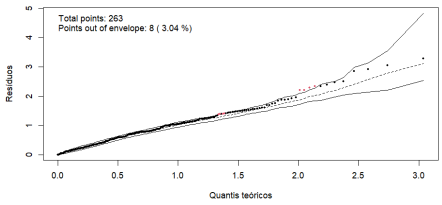
(c) Resíduos Quantílicos.

Os gráficos de resíduos *vs* valores ajustados apresentados na Figura 15, evidenciam um comportamento consideravelmente mais adequado quando comparados aos modelos anteriormente. Fica evidente uma dispersão aproximadamente contante em torno de zero, ficando evidente uma melhora na região dos menores valores ajustados, onde os modelos de Poisson e Binomial Negativo apresentaram maiores distorções. Esses indicati-

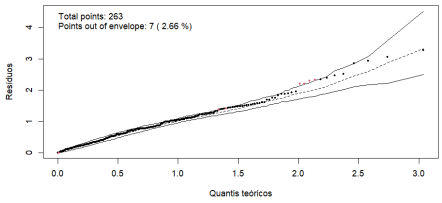
vos sugerem que o modelo ZIP foi capaz de se adequar ao excesso de zeros, reforçando a hipótese da existência de zeros estruturais no processo gerador da amostra.

Em contraste, os gráficos de HNP com envelopes simulados expostos para os resíduos de e Figura 16, demonstram uma melhora expressiva no ajuste do modelo. Os residuos acompanham a linha diagonal de referência, com poucas observações fora dos envelopes de confiança (inferior a 5% dos resíduos). A ausência de desvios indicam que a estrutura de mistura do modelo foi eficaz em acomodar o excesso de zeros presente na amostra. Diante disso, o modelo se mostra possivelmente a escolha mais robusta adequada para o conjunto de dados.

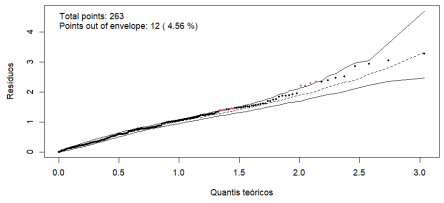
Figura 16 – Gráficos HNP com envelopes simulados - Modelo ZIP.



(a) Resíduos de Pearson.



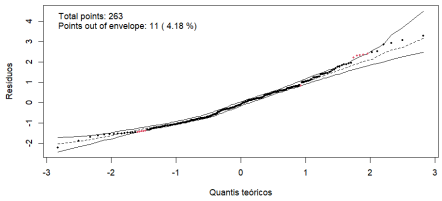
(b) Resíduos de *Deviance*.



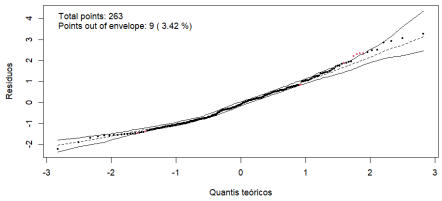
(c) Resíduos Quantílicos.

No que se refere à análise dos gráficos de Q-Q plot com envelopes simulados dos resíduos, exibidos na Figura 17, que comparam os quantis empíricos dos resíduos do modelo ajustado com os quantis teóricos esperados, existe evidencia de um ajuste altamente satisfatório. Observa-se elevada aderência à linha diagonal de referência, bem como uma frequência muito baixa de observações fora dos envelopes de confiança. Esses resultados reforçam a qualidade do ajuste proporcionada pela estrutura de mistura do modelo ZIP ao conjunto de dados.

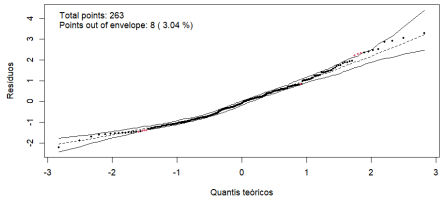
Figura 17 – Gráfico Quantis-Quantis com envelopes simulados - Modelo ZIP.



(a) Resíduos de Pearson.



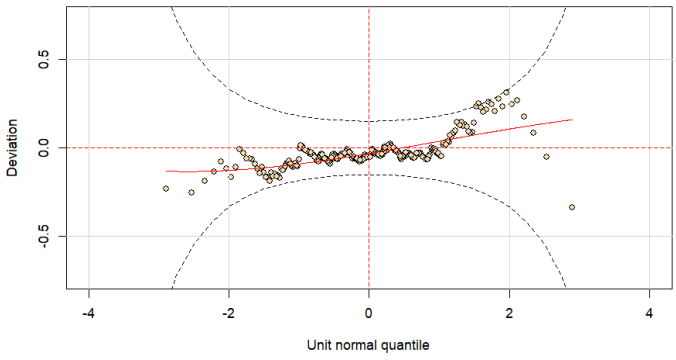
(b) Resíduos de *Deviance*.



(c) Resíduos Quantílicos.

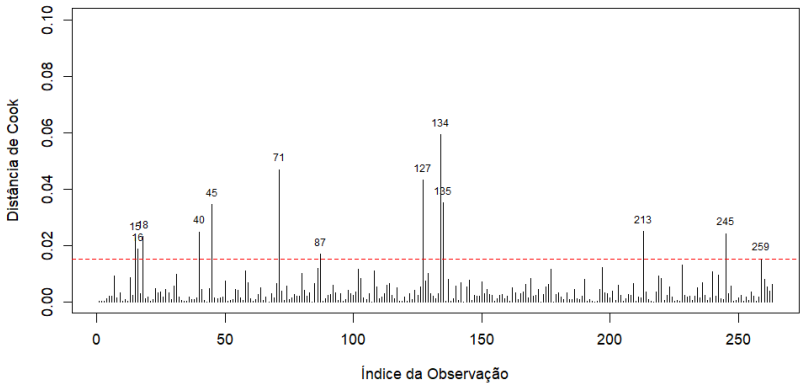
A inspeção do gráfico de *Worm-Plot* exibido na Figura 18, revela um ajuste satisfatório do modelo ZIP aos dados. Observa-se que a grande maioria dos resíduos permanece dentro do envelope de confiança de 95%. Ademais, a linha vermelha apresenta-se próxima ao eixo horizontal, indicando que o modelo foi capaz absorver adequadamente tanto a superdisperssão quanto a presença de excesso de zeros dos dados.

Figura 18 – Gráfico *Worm-Plot* para os resíduos Quantílicos - Modelo ZIP.



Em complemtento, a análise de sensibilidade do modelo foi realizada por meio do gráfico da Distância de Cook, apresentado na Figura 19. Nota-se que, embora alguns pontos ultrapassem o limiar de referência sugerido, com destaque para a observação 71 e 134. As distâncias de Cook assumem valores absolutos bastante baixos. Então, isso confirma a robustez do ajuste do modelo, evidenciando que os resultados não são afetados por um pequeno conjunto de observações discrepantes.

Figura 19 – Gráfico de Distância de Cook para o modelo ZIP.



A partir da análise dos diagnósticos do modelo ZIP, observa-se uma melhora substancial no ajuste em comparação aos modelos de Poisson e Binomial Negativo. Há evidências consistentes de que a utilização de modelos de mistura proporciona ganhos significativos na modelagem de dados de contagem com inflação de zeros, resultando em estimativas mais estáveis e confiáveis. Diante disso, com o objetivo de obter o melhor ajuste possível, procede-se ao ajuste do modelo ZINB, buscando avaliar se a incorporação do parâmetro adicional de dispersão contribui para uma melhoria adicional na qualidade do ajuste.

4.5 Modelo Binomial Negativo Inflacionado de Zeros

Diante de uma possível melhora do ajuste, devido à adição do parâmetro de dispersão, torna-se necessária a adoção de uma abordagem ainda mais flexível e robusta. Nesse contexto, foi utilizado o modelo ZINB, uma ampliação do ZIP, que combina uma estrututa responsável por comtabilizar os zeros estruturais com um processo de contagem baseado na distribuição Binomial Negativo, possibilitando ajuste consistênte para o excesso de zeros e a superdispersão do conjunto de dados.

As variáveis explicativas influenciam a componente de contagem do modelo por meio da função de ligação logarítmica, ao passo que a componente inflacionada de zeros é modelada por meio de uma função de ligação logit. Da mesma forma que nos modelos de Poisson, Binomial Negativo e ZIP, adotou-se a mesma estratégia de modelagem das covariáveis, de modo a garantir comparabilidade entre os ajustes. O modelo foi estimado por meio da função `zeroinfl`() da biblioteca `pscl`, e as estimativas dos coeficientes estão

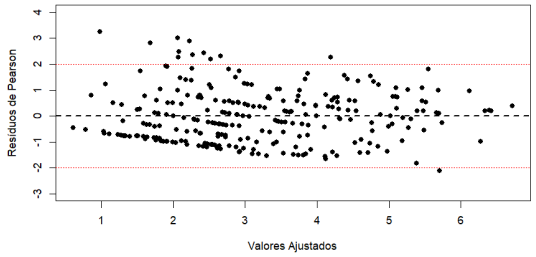
apresentadas na Tabela 7.

Tabela 7 – Resumo da regressão do ZINB.

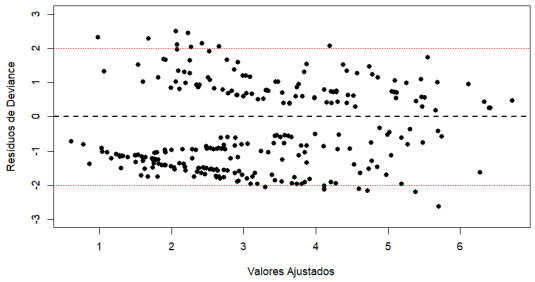
Parâmetros	**Log (Contagem)**			**Logit (Zeros)**		
	$\hat{\beta}$	Erro-Padrão	**Valor-P**	$\hat{\beta}$	Erro-Padrão	**Valor-P**
Intercept	1,384	0,057	$< 0{,}001$	$-1{,}516$	0,282	$< 0{,}001$
DAS	$-0{,}010$	0,003	$< 0{,}001$	0,051	0,018	0,004
AFC	0,009	0,006	0,143	$-0{,}061$	0,029	0,036
SEX	0,003	0,004	0,465	$-0{,}020$	0,017	0,254
Gender	$-0{,}203$	0,079	0,010	$-0{,}061$	0,359	0,865
Infidelity	0,405	0,098	$< 0{,}001$	$-0{,}536$	0,681	0,431
DAS $\times$ Infidelity	0,008	0,006	0,142	$-0{,}081$	0,037	0,030

Na Tabela 7 são apresentados as estimativas, erro-padrão e valor-P para os parâmetros do modelo ZINB. Dessa forma, em paralelo ao modelo ZIP, não identificamos mudanças significativas em relação a seus coeficientes, porém fica claro um aumento em seus erro-padrão, proveniente do parâmetro de dispersão.

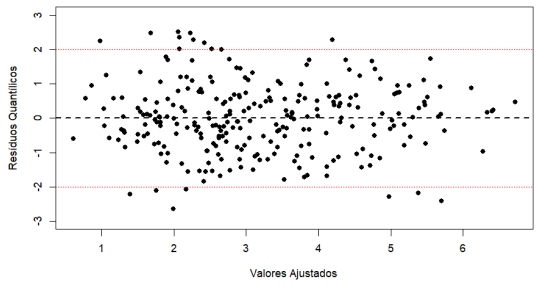
Figura 20 – Gráficos de resíduos *vs* valores ajustados - Modelo ZINB.



(a) Resíduos de Pearson.



(b) Resíduos de *Deviance*.

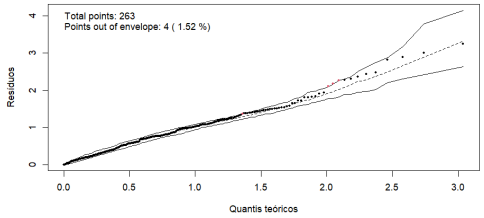


(c) Resíduos Quantílicos.

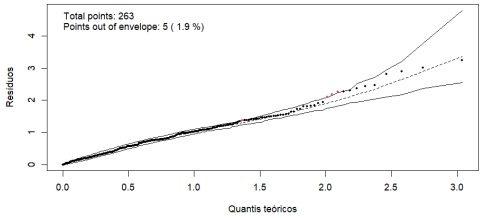
Os gráficos de resíduos *vs* valores ajustados, são apresentada na Figura 20. Ao examinar os resíduos de Pearson (a) e Deviance (b), observam-se padrões, inclinações,

comportamentos esperados e consequência da natureza discreta da variável resposta, não indicando desajuste do modelo. O diagnóstico mais conclusivo está presente nos resíduos quantílicos (c), apresentando uma aleatoriedade dos resíduos, com poucas observações fora dos limites de delimitação. A ausência de estruturas sistemáticas neste gráfico confirma que o modelo ZINB também foi capaz de capturar adequadamente as propriedades de excessos de zeros e superdispersão do conjunto de dados.

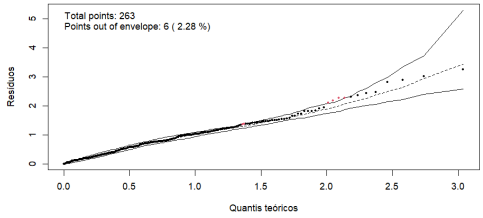
Figura 21 – Gráficos HNP com envelopes simulados - Modelo ZINB.



(a) Resíduos de Pearson.



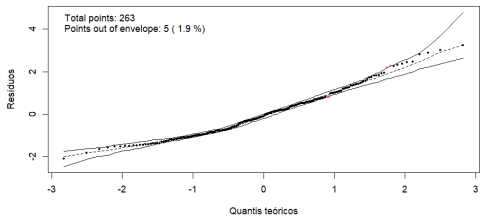
(b) Resíduos de *Deviance*.



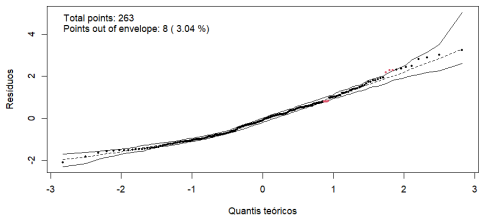
(c) Resíduos Quantílicos.

A validação da bondade de ajuste foi conduzida por meio dos gráficos HNP com envelopes de confiança, exibidos na Figura 21. A inspeção visual revela uma excelente aderência dos resíduos à linha de referência diagonal. Observa-se que a proporção de pontos situados fora dos limites do envelope foram consideravelmente baixos, melhorando as marcas obtidas pelo modelo ZIP. O fato desses percentuais serem inferiores ao nível de significância de 5% sustenta a validade do modelo.

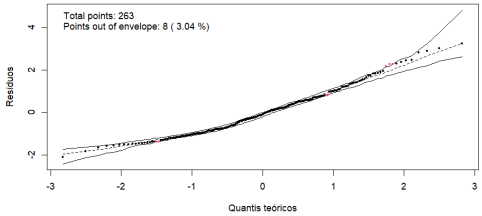
Figura 22 – Gráficos Quantis-Quantis com envelopes simulados - Modelo ZINB.



(a) Resíduos de Pearson.



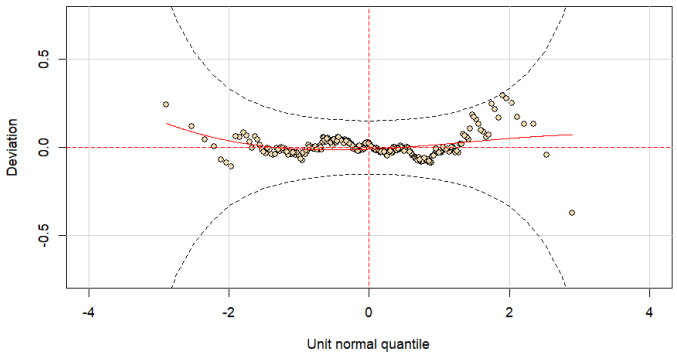
(b) Resíduos de *Deviance*.



(c) Resíduos Quantílicos.

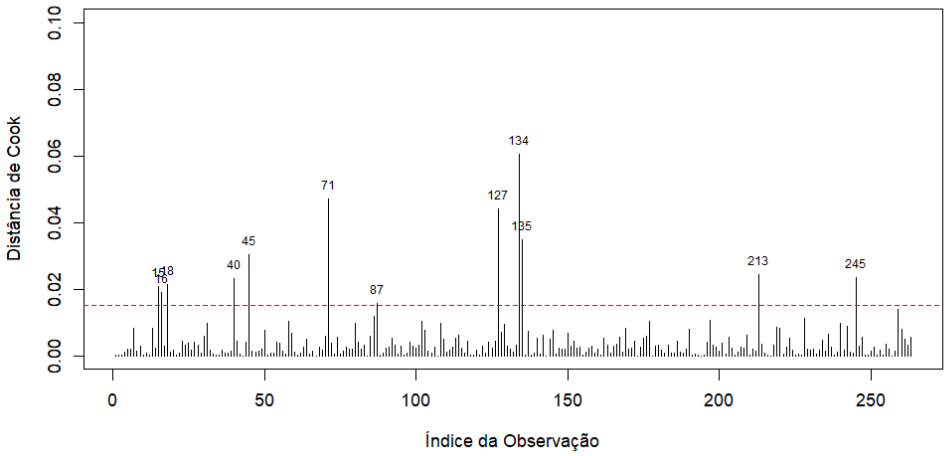
Adicionalmente, a qualidade do ajuste foi corroborada pelos gráficos de Quantil Quantil com envelopes simulados, apresentados na Figura 22. A análise visual demonstra que os resíduos se distribuem linearmente ao longo do eixo característico desse gráfico, sem apresentar observações discrepantes nas caudas. Reforçando a adequação do modelo ZINB para modelar os dados.

Figura 23 – Gráfico *Worm-Plot* para os resíduos Quantílicos - Modelo ZINB.



Por fim, a análise do gráfico *Worm-Plot* para o modelo ZINB, apresentado na Figura 23, demonstra um ajuste de alta qualidade, pode ser identificada uma postura horizontal da linha de suavização, que se sobrepõe quase perfeitamente à linha de referência, o que constitui uma métrica excelente de qualidade de ajuste. Isso evidencia que o modelo de mistura capturou corretamente o excesso de zeros e superdispersão presentes nos dados.

Figura 24 – Gráfico de Distância de Cook para o modelo ZINB.



A análise de sensibilidade do ajuste foi realizada por meio do gráfico da Distância de Cook para o modelo ZINB, apresentado na Figura 24. Esse gráfico permite avaliar o impacto individual de cada observação sobre as estimativas dos parâmetros do modelo. Observa-se que algumas observações apresentam valores acima do patamar de referência indicado pela linha vermelha. Contudo, suas magnitudes absolutas são consideradas baixas, uma vez que os valores máximos não ultrapassam 0,06. Porém, medidas como a remoção das observações possivelmente influentes, podem acarretar na melhora da confiabilidade e na precisão do seu modelo.

Assim, embora os resultados obtidos indiquem um ajuste relativamente satisfatório dos modelos empregados na modelagem dos dados, faz-se necessária a adoção de métodos mais robustos de seleção de modelos, com o objetivo de validar estatisticamente a superioridade de uma distribuição em relação às demais. Para tanto, a análise será complementada pela aplicação de critérios de informação, visando aumentar a eficiência e a precisão na descrição do fenômeno em estudo.

4.6 Resultados Adicionais

A Tabela 8 apresenta os valores dos critérios de AIC e BIC obtidos para cada um dos modelos em estudo Poisson, Binomial Negativo, ZIP e ZINB. Esses critérios ocupam lugar de destaque por serem utilizados para fins de comparação entre modelos, considerando a qualidade do ajuste e a complexidade do modelo em estudo. A tabela permite um panorama comparativo das diferentes medidas empregadas neste trabalho.

Tabela 8 – Critérios de AIC e BIC para os modelos ajustados.

Critérios	Poisson	Binomial Negativo	ZIP	ZINB
AIC	1222,6	1163,2	1111,5	1112,9
BIC	1247,6	1191,8	1161,5	1166,5

Com o objetivo de tornar a escolha do modelo final mais consistente, a Tabela 9 reúne os resultados do teste de Vuong aplicado às comparações entre os modelos Poisson, Binomial Negativo, ZIP e ZINB. O teste de Vuong é amplamente empregado para avaliar a superioridade entre modelos não aninhados, esse teste usa a verossimilhança como critério. Na tabela, são apresentadas as estatísticas Z e os seus respectivos p-valores para as comparações realizadas.

Tabela 9 – Teste de *Vuong* para os modelos ajustados.

Comparações dos modelos	Z	Valor-P
Poisson vs Binomial Negativo	-2,99	0,001
Binomial Negativo vs ZIP	-3,97	$<$0,001
ZIP vs ZINB	-0,36	0,358

O teste de Vuong indicou que o modelo Binomial Negativo se ajusta melhor que o Poisson ($Z = -2,99; p = 0,001$), e que o modelo ZIP supera a Binomial Negativo ($Z = -3,97; p < 0,001$), evidenciando a necessidade de tratar o excesso de zeros. No entanto, a comparação entre ZIP e ZINB não foi significativa ($p = 0,358$). Conclui-se que o modelo ZIP é a escolha mais adequada, pois oferece ajuste equivalente ao ZINB, além de ser mais parcimônioso em relação ao ZINB.

5 CONSIDERAÇÕES FINAIS

Este trabalho foi desenvolvido com o objetivo de apresentar a análise de dados de contagem inflacionados de zeros por meio da aplicação de (MLGs), tais como os modelos de Poisson e Binomial Negativo, em conjunto com abordagens baseadas em estruturas de mistura, como os modelos ZIP e ZINB. Adicionalmente, buscou-se desenvolver e aplicar funções que possibilitem a realização de análises diagnósticas dos modelos apresentados, com base em resíduos, bem como a avaliação de observações influentes por meio da Distância de Cook.

No que se refere à aplicação, verificou-se que os modelos ZIP e ZINB apresentaram desempenho consideravelmente superior na análise dos resíduos, resultado corroborado pela avaliação da Distância de Cook, com menor número de observações fora dos limites de especificação. Ambos os modelos obtiveram os melhores valores nos critérios de informação AIC e BIC. Contudo, o teste de *Vuong* indica que o acréscimo de complexidade proveniente do ZINB não resultou em ganhos significativos na qualidade do ajuste em relação ao modelo ZIP (valor-P = 0,358).

Assim, evidencia-se que abordagens baseadas em estruturas de mistura, como os modelos ZIP e ZINB, são altamente indicadas para a modelagem de dados de contagem inflacionados de zeros, por apresentarem propriedades que permitem acomodar de forma eficiente o excesso de zeros e a superdispersão presentes nos dados. Resultando em estimativas mais consistentes e interpretações mais realistas do evento em estudo.

Ao desenvolver este estudo, observou-se uma quantidade ainda limitada de trabalhos na literatura voltados à área computacional no que se refere a medidas de diagnóstico e análise de resíduos, especialmente para os modelos ZIP e ZINB. Diante disso, como perspectivas para trabalhos futuros, destaca-se a possibilidade de desenvolvimento de uma biblioteca no *software R* que permita a obtenção rápida de resíduos, o cálculo de observações influentes (Cook, 1986), bem como a realização de análises gráficas. Adicionalmente, uma possibilidade seria o desenvolvimento de funções específicas para a avaliação da influência local nos modelos mencionados.

REFERÊNCIAS

AGRESTI, A. **Categorical Data Analysis**. 2. ed. Hoboken, NJ: John Wiley & Sons, 2002.

AKAIKE, H. A new look at the statistical model identification. **Automatic Control, IEEE Transactions on**, v. 19, p. 120–125, 1974.

ATKINS, D. C.; GALLOP, R. J. Rethinking how family researchers model infrequent outcomes: A tutorial on count regression and zero-inflated models. **Journal of Family Psychology**, American Psychological Association, v. 21, n. 4, p. 726–735, 2007.

BUSSAB, W. d. O.; MORETTIN, P. A. **Estatística Básica**. 6. ed. São Paulo: Saraiva, 2009.

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

CARVALHO, F. J. **Modelos lineares generalizados na agronomia: análise de dados binomiais e de contagem, zeros inflacionados e enfoque bayesiano**. Tese (Tese de Doutorado) – Universidade Federal de Uberlândia (UFU), Uberlândia, 2019. Orientadora: Denise Garcia de Santana.

CHRISTENSEN, A.; ATKINS, D. C.; BERNS, S.; WHEELER, J.; BAUCOM, D. H.; SIMPSON, L. E. Traditional versus integrative behavioral couple therapy for significantly and chronically distressed married couples. **Journal of Consulting and Clinical Psychology**, v. 72, n. 2, p. 176–191, 2004.

COOK, R. D. Detection of influential observation in linear regression. **Technometrics**, v. 19, n. 1, p. 15–18, 1977.

COOK, R. D. Assessment of local influence. **Journal of the Royal Statistical Society. Series B (Methodological)**, Wiley, v. 48, n. 2, p. 133–169, 1986.

CORDEIRO, G. M.; DEMÉTRIO, C. G. B. **Modelos Lineares Generalizados e Extensões**. 1. ed. São Paulo: Editora da Universidade de São Paulo (EDUSP), 2008.

CORDEIRO, G. M.; NETO, E. de A. L. **Modelos Paramétricos**. 1. ed. Recife, PE, Brasil: Editora Universitária da UFPE, 2006.

COSTA, J. V.; SILVEIRA, L. V. d. A.; DONALÍSIO, M. R. 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**, Escola Nacional de Saúde Pública Sergio Arouca, Fundação Oswaldo Cruz, v. 32, n. 8, 2016.

COX, D. R.; SNELL, E. J. A general definition of residuals. **Journal of the Royal Statistical Society. Series B (Methodological)**, v. 30, n. 2, p. 248–275, 1968.

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

DUPUY, J.-F. **Statistical Methods for Overdispersed Count Data**. Cham: Springer, 2019. (Springer Series in Statistics).

FUMES, G. **Uso de modelos inflacionados de zeros na análise de questionários de frequência alimentar**. Dissertação (Mestrado) – Universidade Estadual Paulista, Botucatu, SP, Brasil, 2009. Dissertação de Mestrado em Biometria.

GREENWOOD, M.; YULE, G. U. An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents. **Journal of the Royal Statistical Society**, v. 83, n. 2, p. 255–279, 1920.

HILBE, J. M. **Negative Binomial Regression**. 1. ed. Cambridge: Cambridge University Press, 2007.

HILBE, J. M. **Modeling Count Data**. 2. ed. Cambridge: Cambridge University Press, 2014.

HOCKING, R. R. A biometrics invited paper. the analysis and selection of variables in linear regression. **Biometrics**, v. 32, p. 1–49, 1976.

KULLBACK, S.; LEIBLER, R. A. On information and sufficiency. **The Annals of Mathematical Statistics**, v. 22, p. 79–86, 1951.

LAMBERT, D. Zero-inflated poisson regression, with an application to defects in manufacturing. **Technometrics**, v. 34, n. 1, p. 1–14, 1992.

MALLOWS, C. L. Some comments on cp. **Technometrics**, v. 15, p. 661, 1973.

MARDEN, J. I. Positions and qq plots. **Statistical Science**, v. 19, p. 606–614, 2004.

MCCULLAGH, P.; NELDER, J. A. **Generalized Linear Models**. 2. ed. London: Chapman and Hall, 1989. ISBN 978-0-412-31760-6.

MIN, Y.; AGRESTI, A. Random effect models for repeated measures of zero-inflated count data. **Statistics in Medicine**, v. 24, n. 1, p. 1–18, 2005.

MONTMORT, P. R. de. **Essai d'Analyse sur les Jeux de Hazard**. 2. ed. [*S. l.*]: Jacques Quillau, 1713.

MONTOYA, A. G. M. **Inferência e Diagnóstico em Modelos para Dados de Contagem com Excesso de Zeros**. Dissertação (Dissertação de Mestrado) – Universidade Estadual de Campinas, Instituto de Matemática, Estatística e Computação Científica, Campinas, SP, Brasil, 2009.

MYERS, R. H.; MONTGOMERY, D. C.; VINING, G. G.; ROBINSON, T. J. **Generalized Linear Models with Applications in Engineering and the Sciences**. 2. ed. Hoboken, NJ: John Wiley & Sons, 2010.

NELDER, J. A.; WEDDERBURN, R. W. M. Generalized linear models. **Journal of the Royal Statistical Society. Series A (General)**, v. 135, n. 3, p. 370–384, 1972.

PAULA, G. A. **Modelos de Regressão: com Apoio Computacional**. 1. ed. São Paulo: Editora da Universidade de São Paulo (EDUSP), 2013.

POISSON, S.-D. **Recherches sur la Probabilité des Jugements en Matière Criminelle et en Matière Civile**. Paris: Bachelier, 1837.

Posit Software, PBC. **RStudio: Integrated Development Environment for R**. Boston, MA, 2025. Disponível em: https://posit.co/products/open-source/rstudio/.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2025. Disponível em: https://www.R-project.org/.

RAMALHO, J. J. d. S. **Modelos de Regressão para Dados de Contagem**. Dissertação (Dissertação de Mestrado) – Universidade Técnica de Lisboa, Instituto Superior de Economia e Gestão, Lisboa, Portugal, 1996.

RIDOUT, M.; DEMÉTRIO, C. G. B.; HINDE, J. Models for count data with many zeros. **Proceedings of the XIXth International Biometric Conference**, Cape Town, South Africa, p. 179–192, 1998.

RODRIGUES, T. C. V. **Regressão Binomial Negativa Geograficamente Ponderada: Modelando Superdispersão Espacial**. Dissertação (Dissertação de Mestrado) – Universidade de Brasília, Instituto de Ciências Exatas, Departamento de Estatística, Brasília, DF, Brasil, 2012.

SCHWARZ, G. Estimating the dimension of a model. **The Annals of Statistics**, v. 6, p. 461–464, 1978.

SPANIER, G. B. Measuring dyadic adjustment: New scales for assessing the quality of marriage and similar dyads. **Journal of Marriage and Family**, v. 38, n. 1, p. 15–28, 1976.

VUONG, Q. H. Likelihood ratio tests for model selection and non-nested hypotheses. **Econometrica**, The Econometric Society, v. 57, n. 2, p. 307–333, 1989.

WEDDERBURN, R. W. M. Quasi-likelihood functions, generalized linear models, and the gauss–newton method. **Biometrika**, v. 61, n. 3, p. 439–447, 1974.

WOOLDRIDGE, J. M. **Econometric Analysis of Cross Section and Panel Data**. 1. ed. Cambridge, MA: The MIT Press, 2001.

YAU, K. K. W.; WANG, K.; LEE, A. H. Zero-inflated negative binomial mixed regression modeling of over-dispersed count data with extra zeros. **Biometrical Journal**, v. 45, p. 437–452, 2003.

APÊNDICE A – CONDIÇÕES DE REGULARIDADE

No contexto da família exponencial linear, as propriedades utilizadas ao longo deste trabalho, são válidas sob um conjunto de condições usuais de regularidade, tais condições podem ser enunciadas a segue:

Independência do suporte em relação aos parâmetros: O suporte da variável aleatória Y_i, definido por

$$\mathcal{Y} = \{y : f(y; \theta_i, \phi) > 0\},$$

não depende dos parâmetros θ_i e ϕ.

Derivabilidade da log-verossimilhança: A função log-verossimilhança

$$l(\theta_i, \phi; y_i) = \log f(y_i; \theta_i, \phi)$$

é duas vezes diferenciável em relação a θ_i, para quase todo y_i.

Permutação entre derivação e integração: É permitido trocar a ordem entre diferenciação e esperança, isto é,

$$\frac{\partial}{\partial \theta_i}\mathbb{E}\{\ell(\theta_i, \phi; Y_i)\} = \mathbb{E}\left\{\frac{\partial}{\partial \theta_i}\ell(\theta_i, \phi; Y_i)\right\},$$

também para a segunda derivada.

Existência dos momentos relevantes: Existem e são finitos os momentos

$$\mathbb{E}\left|\frac{\partial \ell}{\partial \theta_i}\right|, \quad \mathbb{E}\left|\frac{\partial^2 \ell}{\partial \theta_i^2}\right|, \quad \mathbb{E}\left[\left(\frac{\partial \ell}{\partial \theta_i}\right)^2\right].$$

Identificabilidade do parâmetro: Para quaisquer $\theta_{i1} \neq \theta_{i2}$, tem-se

$$f(y; \theta_{i1}, \phi) \neq f(y; \theta_{i2}, \phi), \quad \text{para algum } y \in \mathcal{Y}.$$

Espaço paramétrico: O verdadeiro valor do parâmetro θ_i pertence ao interior do espaço paramétrico.

APÊNDICE B – CÓDIGOS UTILIZADOS NA APLICAÇÃO

```
#Bibliotecas utilizadas

#Instalar a biblioteca
install.packages("MASS")
install.packages("pscl")
install.packages("car")
install.packages("hnp")
install.packages("DHARMa")
install.packages("statmod")
install.packages("gamlss.ggplots")
install.packages("ggplot2")
install.packages("gamlss")

#Carregar a biblioteca
library(MASS)
library(pscl)
library(car)
library(hnp)
library(statmod)
library(ggplot2)
library(gamlss)
library(DHARMa)
library(gamlss.ggplots)

#Convertendo variáveis numéricas em fatores

Estado_Civil$gender = factor(x = Estado_Civil$gender,
                    levels = 1:2, labels = c("Wife","Husband"))
Estado_Civil$infidelity = factor(x = Estado_Civil$infidelity,
                    levels = 0:1, labels = c("No infidelity","Infidelity"))

#Centralizar os preditores

Estado_Civil$das.c = with(Estado_Civil, das - mean(das))
Estado_Civil$afc.c = with(Estado_Civil, afc - mean(afc))
Estado_Civil$sex.c = with(Estado_Civil, sex - mean(sex))
```

```
with(Estado_Civil, {
  print(contrasts(gender))
  print(contrasts(infidelity))
})

##########Modelos##########

#####Modelo Poisson#####

set.seed(509593)

#Ajustar o Modelo Poisson
Msi.Poisson = glm(msi ~ das.c*infidelity + gender + afc.c + sex.c,
                  data = Estado_Civil, family = "poisson"(link = "log"))

#Summary do modelo Poisson
summary(Msi.Poisson)

#Obtenção do AIC para o modelo Poisson
AIC(Msi.Poisson)

#Obtenção do BIC para o modelo Poisson
BIC(Msi.Poisson)

###Resíduos de Pearson Padronizados para o modelo Poisson###

#Coeficientes estimados
Coef.Poisson = coefficients(Msi.Poisson)

#Matriz X
X = model.matrix(Msi.Poisson)

#Predição do Modelo
Predit.Poisson = exp(X%*%Coef.Poisson)

#Variância do Modelo Poisson
Var.Poisson = exp(X%*%Coef.Poisson)
```

```
#Matriz de pesos
W = diag(Msi.Poisson$weights)

#Matriz de projeção
H = sqrt(W)%*%X%*%solve(t(X)%*%W%*%X)%*%t(X)%*%sqrt(W)

#Diagonal da matriz de projeção
H.Diag = diag(H)

#Resíduos de Pearson Padronizados
Pearson.Padro.Poisson = (msi - Predit.Poisson)/sqrt(Var.Poisson*(1 - H.Diag))

###Gráfico Resíduos versus Valores Ajustados para a Poisson###

plot(x = Predit.Poisson,
     y = Pearson.Padro.Poisson,
     xlab = "Valores Ajustados",
     ylab = "Resíduos de Pearson Padronizados",
     pch = 19,
     col = "black",
     ylim = c(-4, 4.5))

#Linha de referência no zero
abline(h = 0, lty = 2, lwd = 2)

#Linhas de referência em -2 e 2
abline(h = c(-2, 2), col = "red", lty = 3)

###Gráfico Half Normal Plot para a Poisson###

hnp(Msi.Poisson,
    halfnormal = TRUE,
    diagfun = Pearson.Padro.Poisson,
    print.on = TRUE,
    paint.out = TRUE,
    xlab = "Quantis teóricos",
    ylab = "Resíduos",
    pch = 19)
```

```
###Gráfico Quantis-Quantis com envelopes simulados para a Poisson###

hnp(Msi.Poisson,
    halfnormal = FALSE,
    diagfun = Pearson.Padro.Poisson,
    print.on = TRUE,
    paint.out = TRUE,
    xlab = "Quantis teóricos",
    ylab = "Resíduos",
    pch = 19)

######Resíduos de Deviance Padronizados para o modelo Poisson######

Residuos.Deviance.Padro.Poisson = function(Msi.Poisson) {

#Obter os valores ajustados
mu = fitted(Msi.Poisson)

#Obter os valores observados
y = Msi.Poisson$y

#Calcular os resíduos de deviance brutos
sinal = sign(y - mu)

deviance = sinal*sqrt(2*(y*log(y/mu) - (y - mu)))

#Quando y = 0
deviance[y == 0] = -sqrt(2 * mu[y == 0])

#Calcular a matriz de projeção
X = model.matrix(Msi.Poisson)

#Matriz de pesos
W = diag(mu)

sqrtW = sqrt(W)

H = sqrtW%*%X%*%solve(t(X)%*%W%*%X)%*%t(X)%*%sqrtW
```

```
#Obter os elementos diagonais da matri h_ii
hii = diag(H)

#Calcular os resíduos de deviance padronizados
Deviance.Padro = deviance/sqrt(1 - hii)

return(Deviance.Padro)
}

#Calcular resíduos padronizados usando a função
Deviance.Padro = Residuos.Deviance.Padro.Poisson(Msi.Poisson)

###Gráfico Resíduos versus Valores Ajustados###

#Predição do Modelo
Predit.Poisson = fitted(Msi.Poisson)

#Definição dos eixos
plot(x = Predit.Poisson,
     y = Deviance.Padro,
     xlab = "Valores Ajustados",
     ylab = "Resíduos de Deviance Padronizados",
     pch = 19,
     col = "black",
     ylim = c(-4, 4))

#Linha de referência no zero
abline(h = 0, lty = 2, lwd = 2)

#Linhas de referência em -2 e 2
abline(h = c(-2, 2), col = "red", lty = 3)

###Gráfico Half Normal Plot###

hnp(Msi.Poisson,
    halfnormal = TRUE,
    diagfun = Deviance.Padro,
    print.on = TRUE,
    paint.out = TRUE,
```

```
    xlab = "Quantis teóricos",
    ylab = "Resíduos",
    pch = 19)

###Gráfico Quantis-Quantis com envelopes simulados###

hnp(Msi.Poisson,
    halfnormal = FALSE,
    diagfun = Deviance.Padro,
    print.on = TRUE,
    paint.out = TRUE,
    xlab = "Quantis teóricos",
    ylab = "Resíduos",
    pch = 19)

######Resíduos Quantilicos######

set.seed(509593)

Residuos.Quantilicos.Poisson = function(modelo) {

#Obtém resposta observada
y = modelo$y

#Obtém média predita mi da parte Poisson
Mu = fitted(modelo)

#Calcula F(y-1) = P(Y  y-1)
Fy.1 = ppois(y - 1, lambda = Mu)

#Calcula F(y) = P(Y  y)
Fy = ppois(y, lambda = Mu)

#Gera U ~ Uniform(F(y-1), F(y))
U = runif(length(y), min = Fy.1, max = Fy)

#Resíduo quantílico r_q = Phi^{-1}(U)
Residuos.Quantilicos = qnorm(U)
```

```
return(Residuos.Quantilicos)
}

#Calcular resíduos padronizados usando a função
Res.Qua.Poisson = Residuos.Quantilicos.Poisson(Msi.Poisson)

###Gráfico Resíduos versus Valores Ajustados###

#Predição do Modelo
Predit.Poisson = fitted(Msi.Poisson)

# Definição dos eixos
plot(x = Predit.Poisson,
     y = Res.Qua.Poisson,
     xlab = "Valores Ajustados",
     ylab = "Resíduos Quantílicos",
     pch = 19,
     col = "black",
     ylim = c(-4, 4))

#Linha de referência no zero
abline(h = 0, lty = 2, lwd = 2)

#Linhas de referência em -2 e 2
abline(h = c(-2, 2), col = "red", lty = 3)

###Gráfico Half Normal Plot###

hnp(Msi.Poisson,
    diagfun = Res.Qua.Poisson,
    print.on = TRUE,
    paint.out = TRUE,
    xlab = "Quantis teóricos",
    ylab = "Resíduos",
    pch = 19)

###Gráfico Quantis-Quantis com envelopes simulados###

hnp(Msi.Poisson,
```

```
    halfnormal = FALSE,
    diagfun = Res.Qua.Poisson,
    print.on = TRUE,
    paint.out = TRUE,
    xlab = "Quantis teóricos",
    ylab = "Resíduos",
    pch = 19)

###Gráfico Worm-plot###

wp(resid = Res.Qua.Poisson)

#####Modelo Binomial Negativa#####

Msi.BN = glm.nb(msi ~ das.c*infidelity + gender + afc.c + sex.c,
                data = Estado_Civil)


#Sumarry do modelo Binomial Negativa
summary(Msi.BN)

#Obtenção do AIC para o modelo Binomial Negativa
AIC(Msi.BN)

#Obtenção do BIC para o modelo Binomial Negativa
BIC(Msi.BN)

######Resíduos de Pearson Padronizados######

#Coeficientes estimados
Coef.BN = coefficients(Msi.BN)

#Matriz de Desenho
X = model.matrix(Msi.BN)

#Predição do Modelo
Predit.BN = fitted(Msi.BN)

#Parâmetro de dispersão
```

```
theta = Msi.BN$theta

#Variância do Modelo Binomial Negativa
Var.BN = (Predit.BN + Predit.BN^2/theta)

#Matriz de pesos
W = diag(Predit.BN/(1 + Predit.BN/theta))

# Matriz de projeção
H = sqrt(W)%*%X%*%solve(t(X)%*%W%*%X)%*%t(X)%*%sqrt(W)

#diagonal da matriz de projeção
H.Diag = diag(H)

#Resíduos de Pearson Padronizados
Pearson.Padro.BN = (msi - Predit.BN)/sqrt(Var.BN*(1 - H.Diag))

###Gráfico Resíduos versus Valores Ajustados###

#Predição do Modelo
Predit.BN= fitted(Msi.BN)

#Definição dos eixos
plot(x = Predit.BN,
     y = Pearson.Padro.BN,
     xlab = "Valores Ajustados",
     ylab = "Resíduos de Pearson Padronizados",
     pch = 19,
     col = "black",
     ylim = c(-4, 4))

#Linha de referência no zero
abline(h = 0, lty = 2, lwd = 2)

#Linhas de referência em -2 e 2
abline(h = c(-2, 2), col = "red", lty = 3)

###Gráfico Half Normal Plot###
```

```
hnp(Msi.BN,
    halfnormal = TRUE,
    diagfun = Pearson.Padro.BN,
    print.on = TRUE,
    paint.out = TRUE,
    xlab = "Quantis teóricos",
    ylab = "Resíduos",
    pch = 19)

###Gráfico Quantis-Quantis com envelopes simulados###

hnp(Msi.BN,
    halfnormal = FALSE,
    diagfun = Pearson.Padro.BN,
    print.on = TRUE,
    paint.out = TRUE,
    xlab = "Quantis teóricos",
    ylab = "Resíduos",
    pch = 19)

######Resíduos de Deviance Padronizados######

Residuos.Deviance.Padro.BN = function(Modelo.BN){

#Obter os valores ajustados e observados
mu = fitted(Modelo.BN)

#Obter os valores observados
y = Modelo.BN$y

#Obter o parâmetro theta do modelo
theta = Modelo.BN$theta

#Calcular os resíduos de deviance
termo1 = ifelse(y == 0, 0, y * log(y/mu))

termo2 = (y + theta) * log((y + theta) / (mu + theta))

#Cálculo da deviance total para cada ponto
```

```
dev_i = 2 * (termo1 - termo2)

#O sinal do resíduo é baseado na diferença
sinal = sign(y - mu)

#Resíduo de Deviance
deviance = sinal * sqrt(abs(dev_i))

#Calcular a matriz projeção
X = model.matrix(Modelo.BN)

hii = mu/(1 + (mu/theta))

W = diag(hii)

sqrtW = sqrt(W)

#Cálculo da Matriz H
H = sqrtW %*% X %*% solve(t(X) %*% W %*% X) %*% t(X) %*% sqrtW

#Obter os elementos diagonais da matriz h_ii
hii = diag(H)

#Calcular os resíduos de deviance padronizados
Deviance.Padro = deviance / sqrt(1 - hii)

return(Deviance.Padro)
}

#Calcular resíduos padronizados usando a função manual
Deviance.Padro.BN = Residuos.Deviance.Padro.BN(Msi.BN)

###Gráfico Resíduos versus Valores Ajustados###

#Predição do Modelo
Predit.BN = fitted(Msi.BN)

#Definição dos eixos
plot(x = Predit.BN,
```

```
     y = Deviance.Padro.BN,
     xlab = "Valores Ajustados",
     ylab = "Resíduos de Deviance Padronizados",
     pch = 19,
     col = "black",
     ylim = c(-3, 3))

#Linha de referência no zero
abline(h = 0, lty = 2, lwd = 2)

#Linhas de referência em -2 e 2 (intervalo usual de normalidade)
abline(h = c(-2, 2), col = "red", lty = 3)

###Gráfico Half Normal Plot###

hnp(Msi.BN,
    halfnormal = TRUE,
    diagfun = Deviance.Padro.BN,
    print.on = TRUE,
    paint.out = TRUE,
    xlab = "Quantis teóricos",
    ylab = "Resíduos",
    pch = 19)

###Gráfico Quantis-Quantis com envelopes simulados###

hnp(Msi.BN,
    halfnormal = FALSE,
    diagfun = Deviance.Padro.BN,
    print.on = TRUE,
    paint.out = TRUE,
    xlab = "Quantis teóricos",
    ylab = "Resíduos",
    pch = 19)

######Resíduos Quantilicos######

set.seed(509593)
```

```
Residuos.Quantilicos.BN = function(modelo) {

#Resposta observada
y = modelo$y

#Média predita mu
Mu = fitted(modelo)

#Parâmetro de dispersão
theta = modelo$theta

#Probabilidade acumulada F(y-1)
Fy.1 = pnbinom(y - 1, size = theta, mu = Mu)

#Probabilidade acumulada F(y)
Fy = pnbinom(y, size = theta, mu = Mu)

#U ~ Uniform(F(y-1), F(y))
U = runif(length(y), min = Fy.1, max = Fy)

#Resíduo quantílico
Residuos.Quantilicos = qnorm(U)

return(Residuos.Quantilicos)
}

#Obtenção do resíduos quantílicos para o modelo Binomial Negativa
Res.Qua.BN = Residuos.Quantilicos.BN(Msi.BN)

###Gráfico Resíduos versus Valores Ajustados###

#Predição do Modelo
Predit.BN = fitted(Msi.BN)

#Definição dos eixos
plot(x = Predit.BN,
     y = Res.Qua.BN,
     xlab = "Valores Ajustados",
     ylab = "Resíduos Quantílicos",
```

```
     pch = 19,
     col = "black",
     ylim = c(-3, 3))

#Linha de referência no zero
abline(h = 0, lty = 2, lwd = 2)

#Linhas de referência em -2 e 2
abline(h = c(-2, 2), col = "red", lty = 3)

###Gráfico Half Normal Plot###

hnp(Msi.BN,
    halfnormal = TRUE,
    diagfun = Res.Qua.BN ,
    print.on = TRUE,
    paint.out = TRUE,
    xlab = "Quantis teóricos",
    ylab = "Resíduos",
    pch = 19)

###Gráfico Quantis-Quantis com envelopes simulados###

hnp(Msi.BN,
    halfnormal = FALSE,
    diagfun = Res.Qua.BN ,
    print.on = TRUE,
    paint.out = TRUE,
    xlab = "Quantis teóricos",
    ylab = "Resíduos",
    pch = 19)

###Gráfico Worm-plot###

wp(resid = Res.Qua.BN)

#####Modelo ZIP#####

Msi.ZIP = zeroinfl(msi ~ das.c*infidelity + gender + afc.c + sex.c |
```

```
                    das.c*infidelity + gender + afc.c + sex.c,
                  data = Estado_Civil,
                  link = "logit",
                  dist = "poisson",
                  trace = TRUE, EM = TRUE)

#Sumarry do modelo ZIP
summary(Msi.ZIP)

#Obtenção do AIC para o modelo ZIP
AIC(Msi.ZIP)

#Obtenção do BIC para o modelo ZIP
BIC(Msi.ZIP)


Residuos.Pearson.ZIP = function(Modelo.ZIP) {

#Extrair os valores observados
y = Modelo.ZIP$y

#Obter os componentes do modelo Lambda e Pi
Lambda = predict(Modelo.ZIP, type = "count")
Pi = predict(Modelo.ZIP, type = "zero")

#Calcular o Valor Esperado do modelo ZIP
Mu.Esp = (1 - Pi)*Lambda

#Calcular a Variância do modelo ZIP
Var.Esp = Mu.Esp*(1 + Lambda*Pi)

#Calcular o Resíduo de Pearson
Residuos.Pearson = (y - Mu.Esp)/sqrt(Var.Esp)

return(Residuos.Pearson)
}

#Obtendo resíduos de Pearson para o modelo Msi.ZIP
Pearson.ZIP = Residuos.Pearson.ZIP(Msi.ZIP)
```

```
###Gráfico Resíduos versus Valores Ajustados###

#Predição do Modelo
Predit.ZIP = fitted(Msi.ZIP)

#Definição dos eixos
plot(x = Predit.ZIP,
     y = Pearson.ZIP,
     xlab = "Valores Ajustados",
     ylab = "Resíduos de Pearson",
     main = "",
     pch = 19,
     col = "black",
     ylim = c(-4, 4))

#Linha de referência no zero
abline(h = 0, lty = 2, lwd = 2)

#Linhas de referência em -2 e 2
abline(h = c(-2, 2), col = "red", lty = 3)

###Gráfico Half Normal Plot###

hnp(Msi.ZIP,
    halfnormal = TRUE,
    diagfun = Pearson.ZIP,
    print.on = TRUE,
    paint.out = TRUE,
    xlab = "Quantis teóricos",
    ylab = "Resíduos",
    pch = 19)

###Gráfico Quantis-Quantis com envelopes simulados###

hnp(Msi.ZIP,
    halfnormal = FALSE,
    diagfun = Pearson.ZIP,
    print.on = TRUE,
```

```
    paint.out = TRUE,
    xlab = "Quantis teóricos",
    ylab = "Resíduos",
    pch = 19)

######Resíduos de Deviance######

Residuos.Deviance.ZIP = function(Modelo.ZIP) {

#Extrai a variável resposta e parâmetros
y = Modelo.ZIP$y

#Obtém a média da parte Poisson
Mu = predict(Modelo.ZIP, type = "count")

#Obtém a probabilidade de zero inflado
Pi = predict(Modelo.ZIP, type = "zero")

#Cálculo das Probabilidades
P0.ZIP = Pi + ((1 - Pi) * dpois(0, Mu))

Py.ZIP = (1 - Pi) * dpois(y, Mu)

#Log-verossimilhança individual)
Ver.i = ifelse(y == 0, log(P0.ZIP), log(Py.ZIP))

#Log-verossimilhança do modelo saturado
Ver.sat = dpois(y, lambda = y, log = TRUE)

#Cálculo do Sinal
Mu.Global = (1 - Pi) * Mu
Sinal = sign(y - Mu.Global)

#Cálculo do Resíduo de Deviance
Deviance.Component = 2 * (Ver.sat - Ver.i)

#Proteção para garantir que valores muito
#próximos de zero não fiquem negativos
Deviance.Component = ifelse(Deviance.Component < 0, 0, Deviance.Component)
```

```
Residuos.Deviance = Sinal * sqrt(Deviance.Component)

return(Residuos.Deviance)
}

#Obtendo resíduos de Deviance para o modelo Msi.ZIP
Residuos.Dev.ZIP = Residuos.Deviance.ZIP(Msi.ZIP)

###Gráfico Resíduos versus Valores Ajustados###

#Predição do Modelo
Predit.ZIP = fitted(Msi.ZIP)

#Definição dos eixos
plot(x = Predit.ZIP,
     y = Residuos.Dev.ZIP,
     xlab = "Valores Ajustados",
     ylab = "Resíduos de Deviance",
     pch = 19,
     col = "black",
     ylim = c(-3, 3))

#Linha de referência no zero
abline(h = 0, lty = 2, lwd = 2)

#Linhas de referência em -2 e 2
abline(h = c(-2, 2), col = "red", lty = 3)

###Gráfico Half Normal Plot###

hnp(Msi.ZIP,
    halfnormal = TRUE,
    diagfun = Residuos.Dev.ZIP,
    print.on = TRUE,
    paint.out = TRUE,
    xlab = "Quantis teóricos",
    ylab = "Resíduos",
    pch = 19)
```

```
###Gráfico Quantis-Quantis com envelopes simulados###

hnp(Msi.ZIP,
    halfnormal = FALSE,
    diagfun = Residuos.Dev.ZIP,
    print.on = TRUE,
    paint.out = TRUE,
    xlab = "Quantis teóricos",
    ylab = "Resíduos",
    pch = 19)

######Resíduos Quantilicos######

set.seed(509593)

Residuos.Quantilicos.ZIP = function(modelo) {

#Resposta observada
y = modelo$y

#Média Poisson
Mu = predict(modelo, type = "count")

#Probabilidade de zero inflado
Pi0 = predict(modelo, type = "zero")

#Probabilidade acumulada F(y-1)
Fy.1 = numeric(length(y))

for (i in seq_along(y)) {
    if (y[i] == 0) {
      Fy.1[i] = 0
    } else {
      Fy.1[i] = Pi0[i] + (1 - Pi0[i]) * ppois(y[i] - 1, lambda = Mu[i])
    }
  }

#Probabilidade acumulada F(y)
```

```
Fy = Pi0 + (1 - Pi0) * ppois(y, lambda = Mu)

#U ~ Uniform(F(y-1), F(y))
U = runif(length(y), min = Fy.1, max = Fy)

#Resíduos quantílicos
Residuos.Quantilicos = qnorm(U)

return(Residuos.Quantilicos)
}

#Obtendo resíduos Quantílicos para o modelo Msi.ZIP
Res.Qua.ZIP = Residuos.Quantilicos.ZIP(Msi.ZIP)

###Gráfico Resíduos versus Valores Ajustados###

#Predição do Modelo
Predit.ZIP = fitted(Msi.ZIP)

#Definição dos eixos
plot(x = Predit.ZIP,
     y = Res.Qua.ZIP,
     xlab = "Valores Ajustados",
     ylab = "Resíduos de Quantílicos",
     pch = 19,
     col = "black",
     ylim = c(-3, 3))

#Linha de referência no zero
abline(h = 0, lty = 2, lwd = 2)

#Linhas de referência em -2 e 2
abline(h = c(-2, 2), col = "red", lty = 3)

###Gráfico Half Normal Plot###

hnp(Msi.ZIP,
    halfnormal = TRUE,
    diagfun = Res.Qua.ZIP,
```

```
    print.on = TRUE,
    paint.out = TRUE,
    xlab = "Quantis teóricos",
    ylab = "Resíduos",
    pch = 19)

###Gráfico Quantis-Quantis com envelopes simulados###

hnp(Msi.ZIP,
    halfnormal = FALSE,
    diagfun = Res.Qua.ZIP,
    print.on = TRUE,
    paint.out = TRUE,
    xlab = "Quantis teóricos",
    ylab = "Resíduos",
    pch = 19)

###Gráfico Worm-plot###

wp(resid = Res.Qua.ZIP)

#####Modelo ZINB#####

Msi.ZINB = zeroinfl(msi ~ das.c*infidelity + gender + afc.c + sex.c |
                      das.c*infidelity + gender + afc.c + sex.c,
                    data = Estado_Civil,
                    link = "logit",
                    dist = "negbin",
                    trace = TRUE, EM = TRUE)

#Sumarry do modelo ZINB
summary(Msi.ZINB)

#Obtenção do AIC para o modelo ZINB
AIC(Msi.ZINB)

#Obtenção do BIC para o modelo ZINB
BIC(Msi.ZINB)
```

```
######Resíduos de Pearson######

Residuos.Pearson.ZINB = function(Modelo.ZINB) {

#Extrair os valores observados
y = Modelo.ZINB$y

#Obter os componentes do modelo
Lambda = predict(Modelo.ZINB, type = "count")
Pi     = predict(Modelo.ZINB, type = "zero")

#Extrair o parâmetro Theta
Theta = Modelo.ZINB$theta

#Calcular o Valor Esperado
Mu.Esp = (1 - Pi)*Lambda

#Calcular a Variância do modelo ZINB
Var.Esp = Mu.Esp*(1 + Lambda*(Pi + (1 / Theta)))

#Calcular o Resíduo de Pearson
Residuos.Pearson = (y - Mu.Esp)/sqrt(Var.Esp)

return(Residuos.Pearson)
}

#Obtendo resíduos Pearson para o modelo Msi.ZIP
Pearson.ZINB = Residuos.Pearson.ZINB(Msi.ZINB)

###Gráfico Resíduos versus Valores Ajustados###

#Predição do Modelo
Predit.ZINB = fitted(Msi.ZINB)

#Definição dos eixos
plot(x = Predit.ZINB,
     y = Pearson.ZINB,
     xlab = "Valores Ajustados",
     ylab = "Resíduos de Pearson",
```

```
     pch = 19,
     col = "black",
     ylim = c(-3, 4))

#Linha de referência no zero
abline(h = 0, lty = 2, lwd = 2)

#Linhas de referência em -2 e 2
abline(h = c(-2, 2), col = "red", lty = 3)


###Gráfico Half Normal Plot###

hnp(Msi.ZINB,
    halfnormal = TRUE,
    diagfun = Pearson.ZINB,
    print.on = TRUE,
    paint.out = TRUE,
    xlab = "Quantis teóricos",
    ylab = "Resíduos",
    pch = 19)

###Gráfico Quantis-Quantis com envelopes simulados###

hnp(Msi.ZINB,
    halfnormal = FALSE,
    diagfun = Pearson.ZINB,
    print.on = TRUE,
    paint.out = TRUE,
    xlab = "Quantis teóricos",
    ylab = "Resíduos",
    pch = 19)

######Resíduos de deviance######

Residuos.Deviance.ZINB = function(modelo) {

#Extrai a variável resposta observada (y_i)
y = modelo$y
```

```
#Obtém a média (mu_i) prevista da parte NB (contagem)
Mu = predict(modelo, type = "count")

#Obtém a probabilidade de zero-inflado (pi_i)
Pi0 = predict(modelo, type = "zero")

#Extrai o parâmetro de dispersão
Theta = modelo$theta

#Probabilidade prevista de observar zero no modelo ZINB
P0.ZINB = Pi0 + (1 - Pi0) * dnbinom(0, mu = Mu, size = Theta)

#Probabilidade prevista para valores positivos
Py.ZINB = (1 - Pi0) * dnbinom(y, mu = Mu, size = Theta)

#Log-verossimilhança individual do modelo ZINB
Ver.i = ifelse(y == 0, log(P0.ZINB), log(Py.ZINB))

#Log-verossimilhança do modelo saturado
Ver.sat = dnbinom(y, mu = y, size = Theta, log = TRUE)

#Resíduo de deviance
Residuos.Deviance = sign(y - Mu) * sqrt(2 * (Ver.sat - Ver.i))

#Retorna o vetor de resíduos
return(Residuos.Deviance)
}

#Obtendo resíduos de Deviance para o modelo Msi.ZINB
Residuos.Dev.ZINB = Residuos.Deviance.ZINB(Msi.ZINB)

###Gráfico Resíduos versus Valores Ajustados###

#Predição do Modelo
Predit.ZINB = fitted(Msi.ZINB)

#Definição dos eixos
plot(x = Predit.ZINB,
```

```
     y = Residuos.Dev.ZINB,
     xlab = "Valores Ajustados",
     ylab = "Resíduos de Deviance",
     pch = 19,
     col = "black",
     ylim = c(-3, 3))

#Linha de referência no zero
abline(h = 0, lty = 2, lwd = 2)

#Linhas de referência em -2 e 2
abline(h = c(-2, 2), col = "red", lty = 3)

###Gráfico Half Normal Plot###

hnp(Msi.ZINB,
    halfnormal = TRUE,
    diagfun = Residuos.Dev.ZINB,
    print.on = TRUE,
    paint.out = TRUE,
    xlab = "Quantis teóricos",
    ylab = "Resíduos",
    pch = 19)

###Gráfico Quantis-Quantis com envelopes simulados###

hnp(Msi.ZINB,
    halfnormal = FALSE,
    diagfun = Residuos.Dev.ZINB,
    print.on = TRUE,
    paint.out = TRUE,
    xlab = "Quantis teóricos",
    ylab = "Resíduos",
    pch = 19)

######Resíduos Quantilicos######

set.seed(509593)
```

```
Residuos.Quantilicos.ZINB = function(modelo) {

#Resposta observada
y = modelo$y

#Média
Mu = predict(modelo, type = "count")

#Probabilidade de Zero Inflado ()
Pi0 = predict(modelo, type = "zero")

#Parâmetro de dispersão da NB (theta = size)
theta = modelo$theta

#Probabilidade acumulada  F(y-1)
Fy.1 = numeric(length(y))

for (i in seq_along(y)) {
    if (y[i] == 0) {
      Fy.1[i] = 0
    } else {
      Fy.1[i] = Pi0[i] + (1 - Pi0[i]) * pnbinom(y[i] - 1
      , size = theta, mu = Mu[i])
    }
  }

#Probabilidade acumulada F(y)
Fy = Pi0 + (1 - Pi0) * pnbinom(y, size = theta, mu = Mu)

#U ~ Uniform(F(y-1), F(y))
U = runif(length(y), min = Fy.1, max = Fy)


Residuos.Quantilicos = qnorm(U)

return(Residuos.Quantilicos)
}

Residuos.Quant = Residuos.Quantilicos.ZINB(Msi.ZINB)
```

```
###Gráfico Resíduos versus Valores Ajustados###

#Predição do Modelo
Predit.ZINB = fitted(Msi.ZINB)

#Definição dos eixos
plot(x = Predit.ZINB,
     y = Residuos.Quant,
     xlab = "Valores Ajustados",
     ylab = "Resíduos Quantílicos",
     pch = 19,
     col = "black",
     ylim = c(-3, 3))

#Linha de referência no zero
abline(h = 0, lty = 2, lwd = 2)

#Linhas de referência em -2 e 2
abline(h = c(-2, 2), col = "red", lty = 3)

###Gráfico Half Normal Plot###

hnp(Msi.ZINB,
    halfnormal = TRUE,
    diagfun = Residuos.Quant,
    print.on = TRUE,
    paint.out = TRUE,
    xlab = "Quantis teóricos",
    ylab = "Resíduos",
    pch = 19)

###Gráfico Quantis-Quantis com envelopes simulados###

hnp(Msi.ZINB,
    halfnormal = FALSE,
    diagfun = Residuos.Quant,
    print.on = TRUE,
    paint.out = TRUE,
```

```
    xlab = "Quantis teóricos",
    ylab = "Resíduos",
    pch = 19)

###Gráfico Worm-plot###

wp(resid = Residuos.Quant)

############Distância de Cook#############

#Função para calcular a distânica de Cook
Cook.Distance = function(modelo, dados){

  n = nrow(dados)          #Calcular o tamanho da amostra
  beta = coef(modelo)      #Coeficientes originais
  V = vcov(modelo)         #Matriz de covariância original
  p = length(beta)        #Número de parâmetros
  cooks_d = numeric(n)    #Vetor para armazenar as distâncias

  cat("Calculando Distância de Cook\n")
  bp = txtProgressBar(min = 0, max = n, style = 3) #Barra de progresso

  for(i in 1:n) {
    #Tenta ajustar o modelo sem a observação i
    result = tryCatch({
      #Atualiza o modelo removendo a linha i
      model_i = update(modelo, data = dados[-i, ], trace = FALSE)
      #Coeficientes do modelo sem a observação i
      beta_i = coef(model_i)
      #Diferença entre coeficientes
      diff_beta = beta - beta_i
      #Fórmula da Distância de Cook Generalizada
      t(diff_beta)%*%solve(V)%*%diff_beta / p
    }, error = function(e) {
      return(NA)
    })
    cooks_d[i] = result
    setTxtProgressBar(bp, i)
  }
```

```
  close(bp)
  return(cooks_d)
}

#Basta mudar o modelo que você quiser calcular a Distânica de Cook na função
#entre Msi.Poisson, Msi.BN, Msi.ZIP e Msi.ZINB.

#Cálculo usando seu modelo e dados
Cook.Dist = Cook.Distance(Msi.Poisson, Estado_Civil)

#Define um limiar de corte
n = length(msi)
Limite = 4/n

#Visualizar os resultados
    plot(Cook.Dist, type = "h",
    ylim = c(0, 0.1),
    ylab = "Distância de Cook",
    xlab = "Observações")
abline(h = Limite, col = "red", lty = 2)
text(x = which(Cook.Dist > Limite),
     y = Cook.Dist[Cook.Dist > Limite],
     labels = which(Cook.Dist > Limite),
     pos = 3, cex = 0.8)


##########Teste Vuong##########

vuong(Msi.Poisson, Msi.BN)

vuong(Msi.BN, Msi.ZIP)

vuong(Msi.ZIP, Msi.ZINB)
```