Contribuições sobre o envelope simulado na análise de diagnóstico em modelos de regressão

Victor Vinicius Fernandes

Victor Vinicius Fernandes

Contribuições sobre o envelope simulado na análise de diagnóstico em modelos de regressão

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP e ao Departamento de Estatística – DEs-UFSCar, como parte dos requisitos para obtenção do título de Mestre em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística. *VERSÃO REVISADA*

Área de Concentração: Estatística

Orientador: Prof. Dr. Gustavo Henrique de Araujo Pereira

UFSCar/USP – São Carlos
Maio de 2019

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

F363c

Fernandes, Victor Vinicius
Contribuições sobre o envelope simulado na análise de diagnóstico em modelos de regressão / Victor Vinicius Fernandes; orientador Gustavo Henrique de Araujo Pereira. -- São Carlos, 2019.
66 p.

Dissertação (Mestrado - Programa Interinstitucional de Pós-graduação em Estatística) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2019.

1. Envelope simulado. 2. Gráfico de probabilidade normal. 3. Análise de diagnóstico. 4. Modelos de regressão. I. Pereira, Gustavo Henrique de Araujo, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Victor Vinicius Fernandes

Contributions on the simulated envelope for diagnostic analysis in regression models

Master dissertation submitted to the Institute of Mathematics and Computer Sciences – ICMC- USP and to the Department of Statistics – DEs- UFSCar, in partial fulfillment of the requirements for the degree of the Master Interagency Program Graduate in Statistics. REVISED VERSION

Concentration Area: Statistics

Advisor: Prof. Dr. Gustavo Henrique de Araujo Pereira

UFSCar/USP – São Carlos
May 2019

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Ciências Exatas e de Tecnologia
Programa Interinstitucional de Pós-Graduação em Estatística

Folha de Aprovação

Assinaturas dos membros da comissão examinadora que avaliou e aprovou a Defesa de Dissertação de Mestrado do candidato Victor Vinicius Fernandes, realizada em 30/04/2019:

Prof. Dr. Gustavo Henrique de Araujo Pereira
UFSCar

Prof. Dr. Marcio Luis Lanfredi Viola
UFSCar

Prof. Dr. Thiago Maia Magalhães
UFJF

Agradecimentos

Agradeço primeiramente a Deus, pelo dom da vida. Agradeço aos meus familiares, em especial, meus pais, Hélio e Rose, e meus avós, Adão e Dirce, por todo o apoio e incentivo. Agradeço ao meu orientador, Gustavo, pela atenção dedicada ao longo desses anos e por todo o aprendizado. Agradeço aos meus grandes amigos, Dani, Ana Carolina e Paulo, por sempre estarem presentes e torcerem por mim. Por fim, agradeço também aos meus amigos de mestrado, Lucas Cavalaro, Milene, Gabi, Polo, Luiz, Fabi, Juliana, Alex e Lucas Lopes, pelo companheirismo e momentos de estudo. Esta experiência foi incrível!

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES/DS - Demanda Social).

Além disso, esta pesquisa foi desenvolvida com o auxílio dos recursos de HPC disponibilizados pela Superintendência de Tecnologia da Informação da Universidade de São Paulo.

Resumo

Fernandes, V. V. **Contribuições sobre o envelope simulado na análise de diagnóstico em modelos de regressão**.

O envelope simulado é um método da análise de diagnóstico, utilizado para avaliar a veracidade da hipótese referente a distribuição de probabilidade assumida para a variável resposta em um modelo de regressão. Neste trabalho, descrevemos alguns procedimentos para a sua obtenção e, posteriormente, propomos um método para a rejeição do modelo a partir do envelope. No intuito de comparar nosso procedimento com as demais propostas, realizamos um estudo de simulação de Monte Carlo em duas classes de modelos de regressão. Os resultados apontam que o método proposto apresenta boa performance, uma vez que o mesmo fornece taxas estáveis de rejeição do modelo sob a distribuição correta. Já para as demais metodologias, além de possuírem um custo computacional maior, a taxa de rejeição do modelo correto cresce conforme aumenta-se o tamanho amostral. Complementando os resultados, realizamos também a comparação do gráfico de probabilidade normal e meio normal com envelope através de simulações de Monte Carlo. Os estudos sugeriram que, de maneira geral, o gráfico normal demonstrou melhor desempenho, principalmente com a utilização do procedimento proposto. Por fim, aplicamos a dados reais, provenientes da Pesquisa Nacional de Saúde (PNS) de 2013, nosso método de rejeição do modelo e as demais propostas. Constatou-se que para estes dados, nosso método sugeriu uma decisão contrária a fornecida pelos outros procedimentos.

Palavras-chave: *análise de diagnóstico, gráfico de probabilidade normal, modelos de regressão, resíduo, simulação de Monte Carlo*.

Abstract

Fernandes, V. V. **Contributions on the simulated envelope for diagnostic analysis in regression models**.

The simulated envelope is a diagnostic analysis method used to evaluate the hypothesis about the probability distribution assumed for the response variable in a regression model. In this work, we describe some procedures to obtain the simulated envelope and, later, we propose a method to decide if we should reject a model using the envelope. In order to compare our procedure with other proposals, we performed a Monte Carlo simulation study in two classes of regression models. The results indicate that the proposed method presents good performance, since it provides stable rejection rates of the model under the correct distribution. About other methodologies, besides having a higher computational cost, the rejection rate under the correct model increases as the sample size rises. In addition, we also compare the full normal plot and the half normal plot with envelope using Monte Carlo simulations studies. The results suggest that, in general, the full normal plot performs better, especially with the proposed procedure. Finally, we apply our decison method and the other proposals to real data from the National Health Survey (Brazil) of 2013. To these data, our method suggested a different decision from that one provided by the other procedures.

Keywords: diagnostic analysis, Monte Carlo simulation, normal probability plot, regression models, residual.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

As representações gráficas são uma forma de linguagem visual frequentemente presentes em nosso cotidiano. Através delas podemos visualizar a ocorrência de determinados fenômenos de maneira mais clara e informativa (Santos, 2018). Entretanto, somente ao final do século XX, com o avanço dos meios computacionais, que a obtenção de formas gráficas tornou-se relativamente fácil.

Em Estatística, as representações gráficas são preferíveis no diagnóstico de modelos de regressão, uma vez que facilitam a interpretação dos resultados. Dentre as diversas etapas do diagnóstico de um modelo, como o estudo de observações aberrantes, influentes e de alavanca (Paula, 2013), há também a preocupação em avaliar a veracidade da hipótese referente à distribuição de probabilidade assumida para a variável resposta dadas as covariáveis. Para isso, faz-se o uso de resíduos, que sucintamente, podem ser vistos como uma medida de discrepância entre o valor observado e o valor predito pelo modelo (Cook and Weisberg, 1982).

Uma forma gráfica simples de verificar tal hipótese é através do gráfico normal probabilístico (Cordeiro and Demétrio, 2013), em que dispomos na ordenada os resíduos provenientes do modelo e na abcissa os valores esperados das estatísticas de ordem da normal padrão. O intuito é avaliar, por meio dos resíduos, se as observações podem ser consideradas como oriundas de uma amostra aleatória da distribuição que estamos supondo para a variável resposta. Desta maneira, espera-se visualizar um padrão linear dos mesmos, apesar de possíveis flutuações provenientes da amostragem.

Por se tratar de um procedimento informal para verificação da distribuição dos dados, uma grande dificuldade do gráfico de probabilidade normal está em avaliar até que ponto certas irregularidades no padrão linear dos resíduos podem ser consideradas naturais. Além disso, mesmo que o modelo seja correto, os resíduos não são independentes e nem possuem distribuição exatamente normal em amostras finitas (Pereira, 2017).

Uma primeira abordagem para prover um referencial da flutuação dos pontos foi proposta por Ripley (1977) através do uso de simulações. Tal abordagem consiste em replicar amostras simuladas de uma distribuição de referência e sobrepô-las em um único

gráfico probabilístico, com o objetivo de obter uma representação gráfica da variabilidade inerente dos pontos.

Posteriormente, Atkinson (1981) sugere a utilização de uma espécie de banda para a flutuação dos pontos, a qual denominou envelope e tem sido objeto de estudo de diversos pesquisadores. Flack and Flores (1989), por exemplo, investigaram propriedades do método envelope em gráficos normais probabilísticos como instrumento para avaliar o comportamento dos resíduos em modelos de regressão normal. Já Raubertas (1992) examinou as propriedades do envelope como um teste formal de bondade do ajuste, discutindo questões acerca do nível de significância e do poder do teste.

Atualmente, um estudo mais detalhado a respeito do envelope simulado tem se mostrado de grande relevância. Esse fato se verifica através dos inúmeros trabalhos publicados na área de modelos de regressão que fazem uso de tal procedimento. Alguns trabalhos recentes que utilizaram o envelope simulado incluem Barros et al. (2018), Silva Ferreira et al. (2018), Lemonte and Bazán (2018), Galea and Castro (2017) e Santos-Neto et al. (2016).

O principal objetivo deste trabalho é propor, a partir de uma das formas de obtenção do envelope simulado, um procedimento que informe ao usuário se a distribuição de probabilidade assumida para a variável resposta deve ou não ser descartada em um modelo de regressão. Além disso, também é de interesse avaliar qual dos gráficos de probabilidade, normal ou meio normal, provê melhores resultados com a utilização do envelope.

Este trabalho está organizado da seguinte maneira. No Capítulo 2 descrevemos, brevemente, características referentes à inferência sobre os parâmetros nos modelos lineares generalizados (MLG) e, de maneira mais detalhada, aspectos da análise de diagnóstico, abordando especificamente os resíduos. Os modelos de regressão para taxas e proporções são considerados no Capítulo 3, sob a mesma forma de apresentação que os MLG. Já no Capítulo 4, discutimos a respeito dos gráficos de probabilidade, abordamos as formas de construção do envelope simulado e, posteriormente, propomos um método de rejeição do modelo a partir do envelope. O Capítulo 5 é dedicado a estudos de simulação através do método de Monte Carlo com o objetivo de verificar a performance do procedimento de rejeição proposto e avaliar as propriedades dos gráficos de probabilidade normal e meio normal com envelope simulado. No Capítulo 6, aplicamos e comparamos o método de rejeição proposto aos demais procedimentos considerados em dados reais, provenientes da Pesquisa Nacional de Saúde (PNS) de 2013. Por fim, no Capítulo 7 fazemos uma breve recapitulação dos pontos desenvolvidos neste trabalho, apresentamos as principais conclusões obtidas a partir do método de rejeição proposto e do estudo comparativo dos gráficos de probabilidade, além de trabalhos futuros que poderiam ser desenvolvidos.

Capítulo 2

Modelos lineares generalizados

Os modelos de regressão são amplamente utilizados pois estabelecem uma relação funcional entre a média de uma variável de interesse (variável resposta) e variáveis preditoras. Entretanto, as variáveis respostas apresentam certas particularidades, podendo as mesmas serem discretas ou contínuas, limitadas em um intervalo, assimétricas, entre outras características. Contudo, existe na literatura diversas classes de modelos que foram desenvolvidas para acomodar adequadamente tais particularidades. Neste capítulo abordaremos uma classe em específico: os modelos lineares generalizados (MLG).

Iniciamos este capítulo especificando a classe de modelos na Seção 2.1. Em seguida, discutimos na Seção 2.2 um dos métodos para a estimação dos parâmetros do modelo e alguns algoritmos de estimação. Posteriormente, abordamos brevemente na Seção 2.3 aspectos referentes a estimação intervalar e testes de hipótese. Por fim, apresentamos na Seção 2.4 os principais resíduos utilizados na análise de diagnóstico dos MLG. Os assuntos das três primeiras seções deste capítulo são abordados sucintamente, por não serem o interesse principal do trabalho. Já na Seção 2.4, apresentamos a análise de diagnóstico de forma mais detalhada e focada nos resíduos, pois é através deles que obtemos o envelope simulado.

2.1 Especificação do modelo

Os modelos lineares generalizados foram propostos por Nelder and Wedderburn (1972) e permitem a modelagem de variáveis cuja distribuição de probabilidade pertença à família exponencial (FE). As distribuições gama, normal, normal inversa, binomial e poisson são as principais desta classe. Dito isto, definimos os MLG da seguinte forma:

Definição 2.1: Sejam $Y_1, Y_2, \ldots, Y_n$ variáveis aleatórias independentes, cada uma com função de probabilidade ou função densidade de probabilidade expressa na forma da família exponencial dada por

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

para $i = 1, \ldots, n$, em que $E(Y_i) = \mu_i = b'(\theta_i)$ e $VAR(Y_i) = \phi V(\mu_i)$, sendo $V(\mu_i) = b''(\theta_i) = \frac{d\mu_i}{d\theta_i}$ e ϕ o parâmetro de dispersão.

Os modelos lineares generalizados são definidos por $f(y_i; \theta_i, \phi)$ e pelo componente sistemático

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

em que $\eta_i = x_i^T \boldsymbol{\beta}$ é o preditor linear, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^T$ é o vetor de parâmetros desconhecidos, $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$ é formado por constantes que representam os valores das variáveis preditoras associadas a i-ésima unidade amostral e $g(\cdot)$ é a função de ligação estritamente monótona e duplamente diferenciável em relação a μ.

Note que a variância da variável resposta é composta por duas componentes: uma independente da média (parâmetro de dispersão) e a outra variação dependente da mesma (função de variância). Outra observação importante é que a função de variância $V(\mu)$ desempenha um papel importante na família exponencial, uma vez que a mesma caracteriza a distribuição, ou seja, dada a função de variância, tem-se a classe de distribuições correspondente e vice-versa. Na Tabela 2.1 apresentamos as principais distribuições dos MLG pertencentes à família exponencial (Paula, 2013).

Tabela 2.1: Principais distribuições dos MLG e seus termos na forma da FE.

Distribuição	θ	b(θ)	ϕ	V(μ)
Normal	μ	$\theta^2/2$	σ^2	1
Gama	$-1/\mu$	$-log(-\theta)$	σ^2	μ^2
Normal Inversa	$-1/2\mu^2$	$-(-2\theta)^{1/2}$	σ^2	μ^3
Poisson	$log(\mu)$	$\exp(\theta)$	1	μ
Binomial	$log(1/(1-\mu))$	$log(1+\exp(\theta))$	n	$\mu(1-\mu)$

2.2 Estimação dos parâmetros

O procedimento de estimação, comumente utilizado, é baseado no método de máxima verossimilhança (McCullagh and Nelder, 1989). Por conveniência matemática, opta-se por maximizar a função de log verossimilhança ao invés de maximizar a função de verossimilhança, já que ambas são equivalentes em virtude da função logarítmica ser estritamente monótona crescente. Entretanto, na maioria dos casos, a maximização da função de log verossimilhança não possui solução analítica, fazendo-se necessária a utilização de procedimentos numéricos. Dentre os mais conhecidos está o método de Newton-Raphson, que realiza a estimação dos parâmetros por meio da função escore e da matriz hessiana. Há também o método *scoring* de Fisher, que é uma variação do primeiro e consiste em substituir a matriz hessiana pelo seu valor esperado. Maiores detalhes podem ser vistos em Olsson (2002).

Para os modelos lineares generalizados, o procedimento usualmente utilizado para estimar os parâmetros é uma modificação do método *scoring* de Fisher, conhecido como método de mínimos quadrados reponderados iterativamente (Green, 1984).

2.3 Intervalos de confiança e testes de hipótese

Os intervalos de confiança para os parâmetros podem ser obtidos por meio de propriedades assintóticas do estimador de máxima verossimilhança (Sen et al., 2010). No que concerne aos testes de hipóteses, podemos utilizar diferentes estatísticas para a sua realização. A estatística da razão de verossimilhança é usualmente utilizada quando queremos testar hipóteses que envolvem diversos parâmetros. Já a estatística de Wald é a mais simples quando queremos testar hipóteses relativas a um único parâmetro, incluindo o caso de avaliar se este único parâmetro é igual a zero. Além das duas citadas, há também a estatística escore, que é baseada na matriz de informação de Fisher e no vetor escore. Para maiores detalhes das estatísticas de teste, veja Paula (2013). Uma alternativa as demais é a estatística gradiente, que consiste em utilizar o vetor escore e o vetor de parâmetros do modelo (Terrell, 2002). Temos que, sob a hipótese nula em teste, todas essas estatísticas tem distribuição qui-quadrado com q graus de liberdade, sendo q a diferença entre o número de parâmetros sob as hipóteses nula e alternativa.

2.4 Análise de diagnóstico

A análise de diagnóstico é essencial no ajuste de um modelo, pois é através dela que identificamos se existem observações discrepantes, influentes e de alavanca, como também avaliamos se os pressupostos assumidos para a obtenção do modelo são atendidos. Para a verificação das suposições do modelo e identificação de observações discrepantes, faz-se o uso de resíduos, que quantificam a discrepância entre os valores observados e os valores preditos pelo modelo (Cook and Weisberg, 1982). Quando temos apenas uma variável preditora, é fácil realizarmos a análise de diagnóstico. Entretanto, quando o número de covariáveis aumenta, esta tarefa se torna mais difícil. Nesses casos é útil a obtenção da matriz de projeção, discutida a seguir.

Matriz de projeção

A matriz de projeção é assim denominada pois representa a projeção ortogonal do vetor $\mathbf{y}$ no subespaço gerado pelas colunas da matriz $\mathbf{X}$, que é a matriz $n \times p$ formada pelas covariáveis (variáveis preditoras) observadas em cada unidade amostral, sendo as linhas $\mathbf{x_i}^\top = (\mathbf{x_{i1}}, \mathbf{x_{i2}}, \ldots, \mathbf{x_{ip}})$, para $i = 1, \ldots, n$ (Cordeiro and Lima Neto, 2006). Desta forma, temos que a matriz de projeção, representada por $\hat{\mathbf{H}}$, nos MLG é dada por

$$\hat{\mathbf{H}} = \hat{\mathbf{W}}^{1/2}\mathbf{X}(\mathbf{X}^\top\hat{\mathbf{W}}\mathbf{X})^{-1}\mathbf{X}^\top\hat{\mathbf{W}}^{1/2}, \tag{2.2}$$

em que $\hat{\mathbf{W}} = diag\{\hat{w}_1, \hat{w}_2, \ldots, \hat{w}_n\}$, com pesos $\hat{w}_i = \left(\frac{d\hat{\mu}_i}{d\hat{\eta}_i}\right)^2 \frac{1}{V(\hat{\mu}_i)}$. Note que neste caso, $\hat{\mathbf{W}}^{1/2} = diag\{\sqrt{\hat{w}_1}, \sqrt{\hat{w}_2}, \ldots, \sqrt{\hat{w}_n}\}$.

Para a análise de diagnóstico, utilizamos os elementos da diagonal principal da matriz $\hat{\mathbf{H}}$, denotados por $\hat{h}_{ii}$, que consideram tanto a distância que determinada observação i está das demais $n-1$ observações no espaço definido pelas variáveis explicativas, quanto o valor da variável resposta. Além disso, uma das principais funções dos elementos da diagonal da matriz de projeção nos MLG é fazer com que a variância de diferentes resíduos fique aproximadamente constante e próxima de 1.

Com relação aos resíduos, apresentamos quatro formas: o resíduo componente do desvio (Davison et al., 1989), o resíduo de Pearson (McCullagh and Nelder, 1989), o resíduo quantílico (Dunn and Smyth, 1996) e o resíduo quantílico ajustado (Scudilio and Pereira, 2017). As formas residuais serão de fundamental importância para a obtenção dos envelopes simulados, discutidos no Capítulo 4.

Resíduo componente do desvio

A função desvio é dada pela diferença entre a log verossimilhança do modelo saturado, cujo ajuste é feito sem a sumarização dos dados, ou seja, contendo n parâmetros, e a log verossimilhança de um modelo de interesse, contendo p parâmetros, com $p < n$. O i-ésimo componente da função desvio para um MLG é dado por

$$d^2(y_i, \hat{\mu}_i, \hat{\phi}) = \frac{2[y_i(\tilde{\theta}_i - \hat{\theta}_i) + \{b(\hat{\theta}_i) - b(\tilde{\theta}_i)\}]}{\hat{\phi}}, \tag{2.3}$$

em que o termo $\tilde{\theta}_i$ é a estimativa de máxima verossimilhança de θ_i sob o modelo saturado e $\hat{\theta}_i$ é a estimativa de máxima verossimilhança sob o modelo de interesse.

Desta maneira, o resíduo componente do desvio padronizado é expresso por

$$r_i^D = \{\text{sinal}(y_i - \hat{\mu}_i)\}\frac{\sqrt{d^2(y_i, \hat{\mu}_i, \hat{\phi})}}{\sqrt{1 - \hat{h}_{ii}}}, \tag{2.4}$$

em que $\{\text{sinal}(y_i - \hat{\mu}_i)\}$ é 1 se $y_i \geq \hat{\mu}_i$ e -1 se $y_i < \hat{\mu}_i$. O termo $\sqrt{1 - \hat{h}_{ii}}$ é usado para que a variância do resíduo seja aproximadamente constante e próxima de 1.

Resíduo de Pearson

O resíduo de Pearson padronizado é simples e consiste na diferença padronizada entre o valor observado e o valor ajustado pelo modelo. Este resíduo é dado por

$$r_i^P = \frac{(y_i - \hat{\mu}_i)}{\sqrt{\hat{\phi} V(\hat{\mu}_i)(1 - \hat{h}_{ii})}}, \tag{2.5}$$

em que $V(\hat{\mu}_i)$ é a função de variância estimada para a i-ésima observação. Assim como no resíduo componente do desvio, o termo $\sqrt{1 - \hat{h}_{ii}}$ é utilizado no intuito de tornar a variância do resíduo aproximadamente constante e próxima de 1.

Resíduo quantílico

O resíduo quantílico baseia-se no fato de que a função de distribuição acumulada para cada observação, denotada por $F(y_i; \mu_i, \phi)$, possui distribuição uniforme no intervalo unitário. Desta forma, para MLG com variável resposta contínua, o mesmo é definido como

$$r_i^Q = \Phi^{-1}(F(y_i; \hat{\mu}_i, \hat{\phi})), \tag{2.6}$$

em que Φ é a função de distribuição acumulada da normal padrão.

Por meio de simulações via método de Monte Carlo, Feng et al. (2017) compararam o resíduo componente do desvio e o resíduo de Pearson com o resíduo quantílico. Os resultados indicaram que, sob grandes amostras, o resíduo quantílico é melhor para a detecção de falta de ajuste.

Em outro estudo recente, Scudilio and Pereira (2017) avaliaram particularidades dos resíduos anteriormente citados e constataram que o resíduo de Pearson apresenta comportamento assimétrico mesmo em grandes amostras. Já o resíduo componente do desvio, embora seja o mais utilizado e possua maior simetria com relação ao resíduo de Pearson, tanto ele quanto o último apresentam distribuições desconhecidas, mesmo em grandes amostras. Desta forma, o resíduo quantílico é interessante por ter distribuição assintoticamente normal padrão, o que facilita a identificação de pontos discrepantes na análise de diagnóstico.

Resíduo quantílico ajustado

Por meio de estudos de simulação, Scudilio and Pereira (2017) verificaram que o resíduo quantílico dividido pela quantidade $\sqrt{(1-\hat{h}_{ii})}$ apresenta distribuição assintoticamente normal, entretanto, com variância mais próxima de 1 se comparado ao resíduo quantílico convencional. Desta forma, define-se o resíduo quantílico ajustado como

$$r_i^{QH} = \frac{\Phi^{-1}(F(y_i; \hat{\mu}_i, \hat{\phi}))}{\sqrt{(1-\hat{h}_{ii})}}. \tag{2.7}$$

O termo "ajustado" foi usado ao invés de "padronizado" devido a Klar and Meintanis (2012) introduzirem um resíduo denominado resíduo quantílico padronizado para derivar um teste de bondade de ajuste para os modelos lineares generalizados (Scudilio and Pereira , 2017).

Capítulo 3

Modelos de regressão para taxas e proporções

Usualmente há interesse em investigar variáveis contínuas cujo suporte está restrito no intervalo unitário $(0, 1)$, tais como taxas, proporções, porcentagens e afins. Exemplos dessas situações são: a taxa de analfabetismo em um certo país, a proporção de votos para um presidente em exercício concorrendo à reeleição, o percentual do salário gasto com alimentação, dentre outras (Bayes et al., 2012).

Neste capítulo, apresentamos na Seção 3.1 algumas distribuições com suporte no intervalo $(0, 1)$. Já na Seção 3.2, definimos os modelos de regressão baseados nas distribuições vistas na seção anterior para, finalmente na Seção 3.3, abordarmos a análise de diagnóstico nesta classe de modelos.

3.1 Distribuições de probabilidade para taxas e proporções

Diversas distribuições foram desenvolvidas a fim de representar quantidades de interesse que variam no intervalo unitário, dentre elas: a distribuição beta, a distribuição beta retangular (Hahn, 2008), a distribuição simplex (Kieschnick and McCullough, 2003), a distribuição Kumaraswamy (Jones, 2009) e a família de distribuições baseadas na S_B de Johnson (Lemonte and Bazán, 2016), sendo que algumas dessas e outras distribuições com suporte em $(0, 1)$ são descritas em Kotz and Van Dorp (2004). Veremos com mais detalhes apenas as duas primeiras distribuições citadas, uma vez que a distribuição beta retangular será utilizada como concorrente a distribuição beta nos estudos conduzidos via simulação presentes no Capítulo 5. Além disso, optamos pela distribuição beta retangular por ser simples sua geração a partir da distribuição beta.

3.1.1 Distribuição beta

A distribuição beta é frequentemente utilizada para representar variáveis respostas contínuas cujo suporte pertença ao intervalo $(0, 1)$. No entanto, existem na literatura diversas parametrizações para esta distribuição. Utilizaremos aqui a forma paramétrica

em termos de um parâmetro de média μ e um parâmetro de dispersão σ, conforme apresentado em Bayer and Cribari-Neto (2017). Desta maneira, denotamos $Y \sim Beta(\mu, \sigma)$ e temos que sua função densidade de probabilidade é dada por

$$f(y;\mu,\sigma) = \frac{\Gamma\left(\frac{1-\sigma^2}{\sigma^2}\right)}{\Gamma\left(\mu\left(\frac{1-\sigma^2}{\sigma^2}\right)\right)\Gamma\left((1-\mu)\left(\frac{1-\sigma^2}{\sigma^2}\right)\right)} y^{\mu\left(\frac{1-\sigma^2}{\sigma^2}\right)-1}(1-y)^{(1-\mu)\left(\frac{1-\sigma^2}{\sigma^2}\right)-1}, \tag{3.1}$$

em que $0 < y < 1$ e $0 < \sigma < 1$. É possível mostrar que $E(Y) = \mu$ e $VAR(Y) = \sigma^2 V(\mu)$, sendo $V(\mu) = \mu(1-\mu)$ a função de variância. Como já mencionado, σ é um parâmetro de dispersão que varia no intervalo unitário. Essa característica facilita a compreensão de grandeza deste parâmetro, dado que quanto maior o seu valor, maior também será a variância de Y.

Embora a parametrização apresentada não seja a mais comum, a mesma será utilizada na condução dos estudos computacionais (presentes no Capítulo 5) através do pacote *gamlss* (Stasinopoulos et al., 2007) do *software* R, o qual utiliza esta parametrização da distribuição beta.

3.1.2 Distribuição beta retangular

A distribuição beta retangular pode ser vista como a mistura de duas variáveis beta, com a particularidade de que uma delas é $Beta\left(\frac{1}{2}; \frac{1}{\sqrt{3}}\right)$ e portanto, uma $Uniforme(0,1)$, e a outra uma $Beta(\mu, \sigma)$ (Bayes et al., 2012). Denotaremos a distribuição beta retangular como $BR(\mu, \sigma, \lambda)$. Conforme apresentada por Hahn (2008), sua função densidade de probabilidade é expressa por

$$f(y;\mu,\sigma,\lambda) = \lambda + (1-\lambda) f_1(y;\mu,\sigma), \tag{3.2}$$

em que $f_1(y;\mu,\sigma)$ é a função densidade de probabilidade de uma $Beta(\mu,\sigma)$, com $0 < y < 1$, $0 < \sigma < 1$ e $0 < \lambda < 1$ é um parâmetro de mistura. Além disso, temos que

$$E(Y) = \frac{\lambda}{2} + (1-\lambda)\mu \qquad VAR(Y) = (1-\lambda)\sigma^2 V(\mu)\left(1 + \frac{\lambda}{\sigma^2}\right) + \frac{\lambda}{12}(4 - 3\lambda),$$

em que $V(\mu) = \mu(1-\mu)$. Note que obtemos a distribuição uniforme em (3.2) fazendo $\lambda = 1$. Já a distribuição beta é obtida tomando $\lambda = 0$.

3.2 Modelos de regressão para taxas e proporções

Nos modelos de regressão para taxas e proporções, assim como nos modelos lineares generalizados, o interesse em geral está em modelar a média de uma variável resposta em termos de covariáveis. Entretanto, há neste caso a restrição de que $y \in (0,1)$. Apresentamos aqui duas classes de modelos: a regressão beta e a regressão beta retangular.

3.2.1 Modelos de regressão beta

Os modelos de regressão beta foram propostos por Ferrari and Cribari-Neto (2004). Analogamente aos MLG, definimos a regressão beta da seguinte maneira:

Definição 3.1: Sejam $Y_1, Y_2, \ldots, Y_n$ variáveis aleatórias independentes, cada uma com função densidade de probabilidade beta conforme (3.1), com suporte no intervalo $(0,1)$ e o componente sistemático

$$h(\mu_i) = \tau_i,$$

em que $\tau_i = x_i^T\boldsymbol{\beta}$ é o preditor linear, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^T$ é o vetor de parâmetros desconhecidos, $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$ é formado por constantes que representam os valores das variáveis preditoras associadas a i-ésima unidade amostral e $h(\cdot)$ é a função de ligação estritamente monótona e duplamente diferenciável para μ, com domínio em $(0,1)$. Por meio da parametrização utilizada, é possível atribuir diferentes funções de ligação a esta modelagem, tais como a logito, probito, complemento log-log e outras (Bayer, 2011).

A estimação dos parâmetros na regressão beta é feita por máxima verossimilhança através da utilização de métodos numéricos. Assim como nos MLG, os intervalos de confiança para os parâmetros são obtidos através das propriedades assintóticas dos estimadores de máxima verossimilhança (Sen et al., 2010) e os testes de hipótese são realizados por meio das quatro estatísticas apresentadas na Seção 2.3. Os mesmos procedimentos são análogos para a regressão beta retangular, que será discutida a seguir.

3.2.2 Modelos de regressão beta retangular

A regressão beta retangular, propriamente dita, é feita através de uma reparametrização de (3.2) em função da média. Desta forma, se tomarmos $E(Y) = \frac{\lambda}{2} + (1-\lambda)\mu = \nu$, em que Y possui suporte no intervalo $(0,1)$, obtemos que o espaço paramétrico de λ é restrito pelo valor de ν da seguinte forma:

$$0 < \lambda < 1 - |2\nu - 1|.$$

Para obter uma estrutura de regressão mais apropriada da média da distribuição beta retangular, tomamos

$$\nu = \frac{\lambda}{2} + (1-\lambda)\mu \qquad \varphi = \frac{\lambda}{1-(1-\lambda)|2\mu-1|},$$

como a nova parametrização, fazendo com que o espaço paramétrico de ν e φ seja dado pelo retângulo $\{0 \leq \nu \leq 1, 0 \leq \varphi \leq 1\}$. Para maiores detalhes, veja Bayes et al. (2012).

Desta maneira, ν é um parâmetro de média, φ um parâmetro de forma e σ permanece como parâmetro de dispersão. Através da reparametrização feita, a função densidade da distribuição beta retangular é dada por

$$f(y;\nu,\varphi,\sigma) = \varphi(1-|2\nu-1|) + (1-\varphi(1-|2\nu-1|))f_1\left(\frac{\nu-0,5\varphi(1-|2\nu-1|)}{1-\varphi(1-|2\nu-1|)},\sigma\right), \tag{3.3}$$

em que $f_1\left(\frac{\nu-0,5\varphi(1-|2\nu-1|)}{1-\varphi(1-|2\nu-1|)},\sigma\right)$ é a função densidade de uma $Beta\left(\frac{\nu-0,5\varphi(1-|2\nu-1|)}{1-\varphi(1-|2\nu-1|)},\sigma\right)$.

Utilizando a reparametrização em (3.3), definimos a regressão beta retangular da seguinte maneira:

Definição 3.2: Sejam $Y_1, Y_2, \ldots, Y_n$ variáveis aleatórias independentes, cada uma com função densidade de probabilidade beta retangular dada por (3.3), de tal forma que $Y_i \sim BR(\nu_i,\varphi,\sigma)$, e o componente sistemático

$$h(\nu_i) = \tau_i,$$

em que $\tau_i = x_i^T\boldsymbol{\beta}$ é o preditor linear e os demais elementos são definidos tais como apresentados na Subseção 3.2.1.

3.3 Análise de diagnóstico

Os procedimentos para a análise de diagnóstico nos modelos de regressão para taxas e proporções são relativamente recentes. Nesta seção discutiremos os resíduos apenas para os modelos de regressão beta, dado que não utilizaremos ao longo do trabalho os resíduos nos modelos de regressão beta retangular.

Para os resíduos nos modelos de regressão beta, Ferrari and Cribari-Neto (2004) sugeriram uma primeira abordagem por meio do resíduo ordinário padronizado, porém, tal resíduo não é aproximadamente simétrico. A fim de contornar este problema, Espinheira et al. (2008) propuseram outros dois resíduos, denominados resíduo ponderado padronizado 1 e resíduo ponderado padronizado 2. Apresentaremos com mais detalhes os dois últimos resíduos citados, além do resíduo quantílico.

Resíduo ponderado padronizado 1

O resíduo ponderado padronizado 1 (Espinheira et al., 2008) foi proposto com base no algoritmo iterativo *scoring* de Fisher para a estimação dos parâmetros no modelo de regressão beta. Algebricamente, este resíduo é dado por

$$r_i^{W1} = \frac{y_i^\bullet - \hat{\mu}_i^\bullet}{\sqrt{VAR(\hat{y}_i^\bullet)}}, \tag{3.4}$$

em que $y_i^\bullet = log\left(\frac{y_i}{1-y_i}\right)$ e $\hat{\mu}_i^\bullet = \psi\left(\hat{\mu}_i\left(\frac{1-\hat{\phi}^2}{\hat{\phi}^2}\right)\right) - \psi\left((1-\hat{\mu}_i)\left(\frac{1-\hat{\phi}^2}{\hat{\phi}^2}\right)\right)$, sendo $\psi(\cdot)$ a função digama, isto é, $\psi(z) = \frac{d}{dz}log(\Gamma(z))$, para $z > 0$. Além disso, temos que $VAR(\hat{y}_i^\bullet) = \psi'\left(\hat{\mu}_i\left(\frac{1-\hat{\phi}^2}{\hat{\phi}^2}\right)\right) + \psi'\left((1-\hat{\mu}_i)\left(\frac{1-\hat{\phi}^2}{\hat{\phi}^2}\right)\right)$, em que $\psi'(\cdot)$ é a função trigama, dada pela primeira derivada da função digama.

Resíduo ponderado padronizado 2

Com relação ao resíduo ponderado padronizado 2 (Espinheira et al., 2008), o mesmo se trata de uma modificação do resíduo ponderado padronizado 1 e é expresso por

$$r_i^{W2} = \frac{y_i^\bullet - \hat{\mu}_i^\bullet}{\sqrt{VAR(\hat{y}_i^\bullet)(1-\hat{h}_{ii})}}, \tag{3.5}$$

em que $\hat{h}_{ii}$ denota o i-ésimo elemento da diagonal da matriz $\hat{\mathbf{H}}$, que analogamente aos MLG, é definida para a regressão beta como

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

em que $\mathbf{X}$ é a matriz que contém as covariáveis observadas em cada unidade amostral e $\hat{\mathbf{W}} = diag\{\hat{w}_1, \hat{w}_2, \ldots, \hat{w}_n\}$, mas com pesos $\hat{w}_i = VAR(\hat{y}_i^\bullet)\left(\frac{1-\hat{\phi}^2}{\hat{\phi}^2}\right)\left(\frac{1}{(h'(\hat{\mu}_i))^2}\right)$.

Resíduo quantílico

De maneira semelhante ao que foi apresentado na Seção 2.4 para os MLG, definimos na regressão beta o resíduo quantílico com variável resposta contínua como

$$r_i^Q = \Phi^{-1}(F(y_i; \hat{\mu}_i, \hat{\sigma})). \tag{3.6}$$

Recentemente, Pereira (2017) comparou via simulação os resíduos padronizado 1 e 2 conjuntamente com o quantílico. Os resultados sugerem que a distribuição do resíduo quantílico é melhor aproximada pela distribuição normal padrão que os demais na maioria dos cenários avaliados.

Capítulo 4

Envelope simulado

Dentre as diversas etapas da análise de diagnóstico de um modelo, há o interesse em se avaliar a veracidade da hipótese referente à distribuição de probabilidade assumida para a variável resposta dadas as covariáveis. Uma forma gráfica para isso é utilizar o envelope simulado, que se trata da inclusão, em um gráfico de probabilidade, de bandas obtidas por meio de simulações que fornecem um referencial para a flutuação dos pontos.

Neste capítulo, definimos e descrevemos na Seção 4.1 as formas de obtenção dos gráficos de probabilidade. Na Seção 4.2, apresentamos como é feita a construção do envelope conforme Atkinson (1981) para, posteriormente, abordarmos uma forma alternativa e apresentarmos modificações sugeridas por Flack and Flores (1989) e Lee and Rhee (2001). Ainda nesta seção, discutimos brevemente propostas de utilização do envelope simulado como um teste formal de bondade de ajuste para, em seguida, com base nos procedimentos já existentes de concepção do envelope, propomos na Seção 4.3 um método que, a partir do envelope, sugere se a distribuição suposta para a variável resposta deve ou não ser descartada.

4.1 Gráficos de probabilidade

O gráfico de probabilidade é uma técnica gráfica informal utilizada para comparar duas distribuições de probabilidade. Para isso, existem dois tipos: o gráfico quantil (*QQ-plot*) e o gráfico da distribuição acumulada (*PP-plot*). O primeiro é comumente utilizado e consiste em dispor no gráfico os quantis de duas variáveis de interesse, de tal forma que se ambas forem identicamente distribuídas, espera-se visualizar um padrão linear dos pontos. O gráfico da distribuição acumulada possui o mesmo objetivo do gráfico quantil, entretanto, ao invés do quantil das variáveis, utiliza-se a função de distribuição acumulada para realizar a comparação (Gan and Koehler, 1990).

Neste trabalho abordaremos os gráficos quantis, utilizando de maneira mais específica os valores esperados das estatísticas de ordem da normal padrão no eixo da abcissa. Tais gráficos, sob essa configuração, são denominados gráficos de probabilidade normal, que ao longo das últimas décadas foram amplamente estudados. Na década de 50,

Chernoff and Lieberman (1954) sugeriram as primeiras diretrizes de utilização do gráfico normal de probabilidade, dado que, até aquele momento, nenhum material literário havia abordado o procedimento de construção deste gráfico de maneira específica. Posteriormente, Daniel (1959) ao avaliar a significância dos efeitos em experimentos fatoriais 2^k não replicados, propôs e fez uso do gráfico meio normal de probabilidade, o qual consiste em obter os valores absolutos das estatísticas de ordem da normal padrão.

Há outros trabalhos relevantes a respeito dos gráficos de probabilidade que podem ser mencionados. Wilk and Gnanadesikan (1968), por exemplo, descreveram e discutiram técnicas gráficas baseadas nos gráficos de probabilidade na avaliação de amostras unidimensionais, tanto utilizando os dados observados quanto quantidades resultantes da análise dos mesmos. Já Zahn (1975), em um estudo mais prático, propôs modificações na versão original do gráfico de probabilidade meio normal e avaliou por meio de simulações seu comportamento em experimentos fatoriais sob diferentes condições. McCullagh and Nelder (1989) sugeriram o uso do gráfico meio normal probabilístico para medidas não negativas, como é o caso dos elementos da diagonal da matriz de projeção e medidas de influência, como, por exemplo, a distância de Cook, e do gráfico normal probabilístico para quantidades tais como os resíduos. Recentemente, Nóbrega (2010) realizou um estudo comparativo dos gráficos de probabilidade normal e meio normal, a fim de verificar qual deles apresenta melhores propriedades de uso em experimentos fatoriais não replicados.

Feita a introdução aos gráficos de probabilidade e uma breve recapitulação bibliográfica, descreveremos com mais detalhes a obtenção do gráfico normal e meio normal de probabilidade dispondo no eixo da ordenada os resíduos provenientes de um modelo de regressão. Tal especificação é importante, pois será através dela que desenvolveremos os conceitos presentes nas seções posteriores.

Para a obtenção do gráfico de probabilidade normal, considere t_i, para $i = 1, \ldots, n$, um resíduo qualquer. Colocamos no gráfico os pontos $(E(Z_{[i]}), t_{[i]})$, em que $[i]$ representa o i-ésimo menor valor e $E(Z_{[i]})$ se refere aos valores esperados das estatísticas de ordem da normal padrão. Temos que

$$E(Z_{[i]}) \cong \Phi^{-1}\left(\frac{i - 3/8}{n + 1/4}\right), \tag{4.1}$$

em que Φ é a função de distribuição acumulada da normal padrão. Tal aproximação foi primeiramente proposta por Blom (1958) ao investigar a plotagem de pontos em gráficos de probabilidade normal.

Já para o gráfico meio normal probabilístico, tomamos o valor absoluto de t_i, com $i = 1, \ldots, n$. Dispomos no gráfico os pontos $(E(|Z_{[i]}|), |t_{[i]}|)$, em que $[i]$ representa o i-ésimo menor valor e $E(|Z_{[i]}|)$ agora se refere aos valores esperados de $|Z_{[i]}|$, que obtemos fazendo

$$E(|Z_{[i]}|) \cong \Phi^{-1}\left(\frac{n + i + 1/2}{2n + 9/8}\right), \tag{4.2}$$

sendo Φ também a função de distribuição acumulada da normal padrão.

A Figura 4.1 apresenta uma ilustração dos gráficos de probabilidade normal e meio normal para dados simulados. Os dados foram gerados de uma amostra de tamanho 200 da distribuição gama, cujo intervalo de variação teórico da média está compreendido em (20,086; 403,429) e o parâmetro de dispersão é $\phi = 0,25$. Em seguida, ajustamos a eles um modelo de regressão gama e obtivemos o resíduo componente do desvio. Com isso, torna-se possível avaliar se as observações, por meio dos resíduos, podem ser consideradas como oriundas de uma amostra aleatória da distribuição que supomos para a variável resposta. Como já mencionado, se o modelo for adequado, espera-se visualizar um padrão linear dos pontos, apesar de possíveis flutuações provenientes da amostragem.

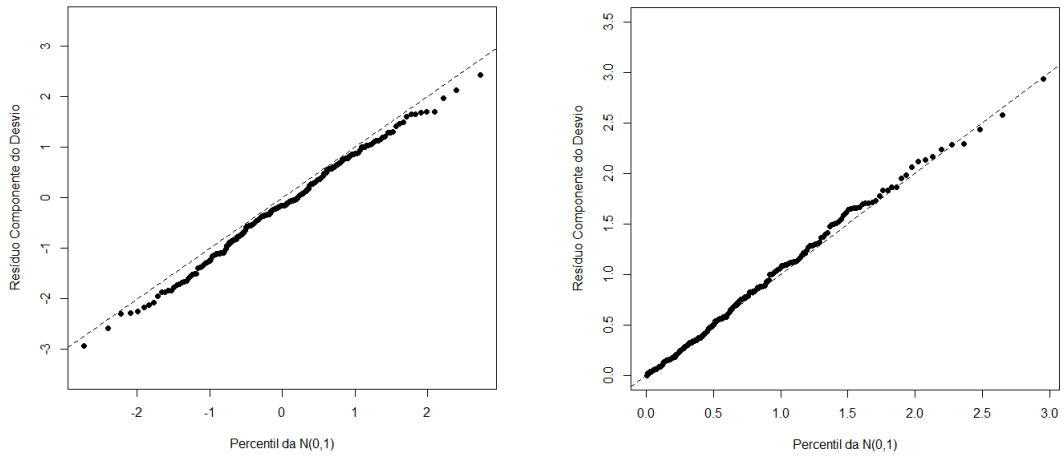

Figura 4.1: Gráficos de probabilidade normal (esquerda) e meio normal (direita) para dados gerados assumindo um modelo de regressão gama.

Por se tratar de um método informal para verificar a distribuição da variável resposta, uma grande dificuldade dos gráficos de probabilidade normal e meio normal está em avaliar até que ponto certas irregularidades no padrão linear dos resíduos podem ser consideradas naturais. Além disso, mesmo que o modelo seja correto, os resíduos não são independentes e nem possuem distribuição exatamente normal padrão em amostras finitas (Pereira, 2017).

Veremos a seguir algumas propostas desenvolvidas para a construção de bandas obtidas através de simulação para a flutuação dos pontos em gráficos de probabilidade.

4.2 Métodos de construção do envelope simulado

Uma primeira abordagem para prover um referencial para a oscilação dos pontos foi proposta por Ripley (1977). Ela consiste em replicar amostras simuladas de uma distribuição de referência e sobrepô-las em um único gráfico probabilístico, com o objetivo de fornecer uma representação gráfica da variabilidade inerente dos pontos. Porém, a proposta que passou a ser comumente utilizada foi a construção de um envelope simulado conforme descrito na próxima subseção.

4.2.1 Envelope simulado: Atkinson

Atkinson (1981) sugeriu a utilização de uma espécie de banda para a flutuação dos pontos, a qual denominou envelope e que tem sido utilizado por diversos pesquisadores. Este procedimento consiste em determinar os limites do envelope realizando os seguintes passos:

1) Geramos n observações $y^*_{i,1}$ considerando que o modelo ajustado é verdadeiro e guardamos esses valores em $\mathbf{y}^*_1 = (y^*_{1,1}, \ldots, y^*_{n,1})^\top$.

2) Ajustamos $\mathbf{y}^*_1$ usando a matriz de variáveis preditoras $\mathbf{X}$ e obtemos os resíduos $t_{i,1}$, para $i = 1, 2, \ldots, n$.

3) Repetimos os passos 1) e 2) m vezes. Logo, teremos $t_{i,j}$, para $i = 1, 2, \ldots, n$ e $j = 1, 2, \ldots, m$, conforme apresentado na Tabela 4.1.

Tabela 4.1: Resíduos originais e simulados.

Resíduo original	Réplica 1	Réplica 2	⋯	Réplica $m-1$	Réplica m
t_1	$t_{1,1}$	$t_{1,2}$	$\ldots$	$t_{1,m-1}$	$t_{1,m}$
t_2	$t_{2,1}$	$t_{2,2}$	$\ldots$	$t_{2,m-1}$	$t_{2,m}$
t_3	$t_{3,1}$	$t_{3,2}$	$\ldots$	$t_{3,m-1}$	$t_{3,m}$
t_4	$t_{4,1}$	$t_{4,2}$	$\ldots$	$t_{4,m-1}$	$t_{4,m}$
t_5	$t_{5,1}$	$t_{5,2}$	$\ldots$	$t_{5,m-1}$	$t_{5,m}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
t_{n-1}	$t_{n-1,1}$	$t_{n-1,2}$	$\ldots$	$t_{n-1,m-1}$	$t_{n-1,m}$
t_n	$t_{n,1}$	$t_{n,2}$	$\ldots$	$t_{n,m-1}$	$t_{n,m}$

4) Colocamos cada grupo de n resíduos em ordem crescente, obtendo assim $t_{[i],j}$, para $i = 1, 2, \ldots, n$ e $j = 1, 2, \ldots, m$, conforme apresentamos na Tabela 4.2.

Tabela 4.2: Resíduos originais e simulados ordenados de forma crescente.

4) ⇓	⇓	⇓		⇓	⇓
Resíduo original	Réplica 1	Réplica 2	⋯	Réplica $m-1$	Réplica m
$t_{[1]}$	$t_{[1],1}$	$t_{[1],2}$	$\ldots$	$t_{[1],m-1}$	$t_{[1],m}$
$t_{[2]}$	$t_{[2],1}$	$t_{[2],2}$	$\ldots$	$t_{[2],m-1}$	$t_{[2],m}$
$t_{[3]}$	$t_{[3],1}$	$t_{[3],2}$	$\ldots$	$t_{[3],m-1}$	$t_{[3],m}$
$t_{[4]}$	$t_{[4],1}$	$t_{[4],2}$	$\ldots$	$t_{[4],m-1}$	$t_{[4],m}$
$t_{[5]}$	$t_{[5],1}$	$t_{[5],2}$	$\ldots$	$t_{[5],m-1}$	$t_{[5],m}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$t_{[n-1]}$	$t_{[n-1],1}$	$t_{[n-1],2}$	$\ldots$	$t_{[n-1],m-1}$	$t_{[n-1],m}$
$t_{[n]}$	$t_{[n],1}$	$t_{[n],2}$	$\ldots$	$t_{[n],m-1}$	$t_{[n],m}$

5) Obtemos os limites $t_{[i],[1]} = \min_j(t_{[i],j})$ e $t_{[i],[m]} = \max_j(t_{[i],j})$. Assim, os limites correspondentes a $t_{[i]}$ serão dados por $t_{[i],[1]}$ e $t_{[i],[m]}$, como presente na Tabela 4.3.

Tabela 4.3: Obtendo os limites correspondentes de cada resíduo no envelope (Atkinson).

	Réplica 1	Réplica 2	...	Réplica $m-1$	Réplica m	Mínimo	Resíduo Original	Máximo
5) $\Rightarrow$	$t_{[1],1}$	$t_{[1],2}$	$\cdots$	$t_{[1],m-1}$	$t_{[1],m}$	$t_{[1],[1]}$	$t_{[1]}$	$t_{[1],[m]}$
$\Rightarrow$	$t_{[2],1}$	$t_{[2],2}$	$\cdots$	$t_{[2],m-1}$	$t_{[2],m}$	$t_{[2],[1]}$	$t_{[2]}$	$t_{[2],[m]}$
$\Rightarrow$	$t_{[3],1}$	$t_{[3],2}$	$\cdots$	$t_{[3],m-1}$	$t_{[3],m}$	$t_{[3],[1]}$	$t_{[3]}$	$t_{[3],[m]}$
$\Rightarrow$	$t_{[4],1}$	$t_{[4],2}$	$\cdots$	$t_{[4],m-1}$	$t_{[4],m}$	$t_{[4],[1]}$	$t_{[4]}$	$t_{[4],[m]}$
$\Rightarrow$	$t_{[5],1}$	$t_{[5],2}$	$\cdots$	$t_{[5],m-1}$	$t_{[5],m}$	$t_{[5],[1]}$	$t_{[5]}$	$t_{[5],[m]}$
	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\Rightarrow$	$t_{[n-1],1}$	$t_{[n-1],2}$	$\cdots$	$t_{[n-1],m-1}$	$t_{[n-1],m}$	$t_{[n-1],[1]}$	$t_{[n-1]}$	$t_{[n-1],[m]}$
$\Rightarrow$	$t_{[n],1}$	$t_{[n],2}$	$\cdots$	$t_{[n],m-1}$	$t_{[n],m}$	$t_{[n],[1]}$	$t_{[n]}$	$t_{[n],[m]}$

O objetivo desse procedimento de simulação é fornecer amostras de resíduos com mesma estrutura de covariância do modelo ajustado (Atkinson, 1981). Além disso, a sugestão de Atkinson é usar $m = 19$, de modo que a probabilidade do maior resíduo em um envelope particular exceder o limite superior seja em torno de 1/20 (Paula, 2013). Analogamente, se considerarmos tal ocorrência para o limite inferior, podemos supor que a probabilidade do maior resíduo ser inferior a esse limite também seja por volta de 1/20, resultando em uma probabilidade aproximada de 0,1 do maior resíduo não estar entre as bandas do envelope.

4.2.2 Envelope simulado: Everitt/Paula

Alternativamente, podemos utilizar m consideravelmente maior e usar como limites inferior e superior não o mínimo e o máximo, mas valores tais que produzam uma espécie de banda de confiança. Uma abordagem segundo essa perspectiva foi proposta por Everitt (1994), que sugeriu utilizar $m = 100$ e determinar as bandas do envelope atribuindo ao limite inferior e ao limite superior os vetores $t_{[i],[5]}$ e $t_{[i],[95]}$, respectivamente, para $i = 1, \ldots, n$. Posteriormente, Paula (2013) propôs uma modificação nesta abordagem, mantendo $m = 100$, mas definindo o limite inferior como a média dos vetores $t_{[i],[2]}$ e $t_{[i],[3]}$, e o limite superior como a média dos vetores $t_{[i],[97]}$ e $t_{[i],[98]}$, para $i = 1, \ldots, n$. Neste trabalho, faremos uso de $m = 99$ e determinaremos os limites do envelope segundo a modificação de Paula. Sendo assim, referenciaremos este procedimento como Everitt/Paula.

4.2.3 Envelope simulado: Flack e Flores

Por meio de estudos mais detalhados em modelos de regressão normal, Flack and Flores (1989) constataram problemas com o procedimento de construção do envelope de Atkinson. O principal deles é o fato do método depender de estatísticas de ordem extremas para a definição dos limites. Isso torna o método suscetível a criação de envelopes bastante variáveis em virtude do aparecimento de valores discrepantes nos resíduos obtidos a partir dos dados simulados.

Devido a esse motivo, Flack e Flores introduziram um outro procedimento para estipular as bandas do envelope segundo uma regra resistente a *outliers*, proposta por Hoaglin et al. (1986) para dados univariados. Basicamente, a primeira diferença em relação ao método de Atkinson está no passo 5, agora definido como

5) Obtemos os limites do envelope fazendo

$$l_i = F_i^L - k\{F_i^U - F_i^L\},$$
$$u_i = F_i^U + k\{F_i^U - F_i^L\},$$

em que F_i^L e F_i^U, para $i = 1, \ldots, n$, são os percentis 25 e 75 da m-ésima estatística de ordem, ou seja, $t_{[i],[0,25m]}$ e $t_{[i],[0,75m]}$, respectivamente, e k é uma constante que pode ser escolhida apropriadamente.

Além disso, considere $\mathbf{l}_1$ e $\mathbf{u}_1$ os vetores que contêm os limites inferior e superior, respectivamente, de uma única réplica do envelope simulado, conforme Tabela 4.4. Repetimos este procedimento 1000 vezes, de tal forma a obter os vetores $\{\mathbf{l}_1, \mathbf{l}_2, \mathbf{l}_3, \ldots, \mathbf{l}_{999}, \mathbf{l}_{1000}\}$ e $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3, \ldots, \mathbf{u}_{999}, \mathbf{u}_{1000}\}$. O envelope final é dado pela média de todos os limites inferiores ($\mathbf{l}^*$) e superiores ($\mathbf{u}^*$) provenientes dos 1000 passos replicados.

Tabela 4.4: Obtendo os limites correspondentes de cada resíduo no envelope (Flack e Flores).

4)	⇓ Réplica 1	⇓ Réplica 2	...	⇓ Réplica $m-1$	⇓ Réplica m	$\mathbf{l}_1$	Resíduo Original	$\mathbf{u}_1$
5) ⇒	$t_{[1],1}$	$t_{[1],2}$	$\ldots$	$t_{[1],m-1}$	$t_{[1],m}$	l_1	$t_{[1]}$	u_1
⇒	$t_{[2],1}$	$t_{[2],2}$	$\ldots$	$t_{[2],m-1}$	$t_{[2],m}$	l_2	$t_{[2]}$	u_2
⇒	$t_{[3],1}$	$t_{[3],2}$	$\ldots$	$t_{[3],m-1}$	$t_{[3],m}$	l_3	$t_{[3]}$	u_3
⇒	$t_{[4],1}$	$t_{[4],2}$	$\ldots$	$t_{[4],m-1}$	$t_{[4],m}$	l_4	$t_{[4]}$	u_4
⇒	$t_{[5],1}$	$t_{[5],2}$	$\ldots$	$t_{[5],m-1}$	$t_{[5],m}$	l_5	$t_{[5]}$	u_5
	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
⇒	$t_{[n-1],1}$	$t_{[n-1],2}$	$\ldots$	$t_{[n-1],m-1}$	$t_{[n-1],m}$	l_{n-1}	$t_{[n-1]}$	u_{n-1}
⇒	$t_{[n],1}$	$t_{[n],2}$	$\ldots$	$t_{[n],m-1}$	$t_{[n],m}$	l_n	$t_{[n]}$	u_n

Com relação a escolha de k, Flack and Flores (1989) sugeriram a utilização de duas quantidades específicas. Uma delas é $k = 1{,}5$, que segundo Hoaglin et al. (1986) é uma boa regra para sinalizar *outliers* em amostras de tamanho moderado no caso univariado. O outro valor considerado é $k = 2{,}25 - 3{,}6/m$, que também no caso univariado representa uma probabilidade aproximada de 5% de uma amostra normal com $m < 100$ possuir um ou mais pontos fora do envelope (Hoaglin and Iglewicz, 1987). Como Flack e Flores utilizam $m = 19$ para a obtenção do envelope, temos para o segundo valor $k \approx 2{,}06$.

4.2.4 Envelope simulado: Lee e Rhee

Após as formas de construção propostas por Atkinson e Flack e Flores, Lee and Rhee (2001) sugeriram um método que combina os métodos anteriores. Os 4 primeiros passos são semelhantes aos do método de Atkinson, mas considerando $m = 2000$. O quinto passo é feito como no procedimento de Flack e Flores, também escolhendo um k apropriadamente.

Lee and Rhee (2001) justificaram o uso dessa nova forma de envelope pelo fato de sua construção não depender de estatísticas de ordem extrema, como ocorre segundo Atkinson, e o procedimento não ser computacionalmente intensivo, como acontece com o método de Flack e Flores. Outro ponto de destaque do método de Lee e Rhee em relação ao procedimento de Flack e Flores é a quantidade de modelos ajustados. Enquanto que para o último são necessários 19000 ajustes para a obtenção do envelope, no método de Lee e Rhee são feitos apenas 2000 modelos, o que reduz consideravelmente o tempo de execução computacional.

Além do procedimento descrito, Lee e Rhee sugeriram uma outra abordagem para estipular as bandas do envelope. A diferença para a versão já apresentada está no passo 5, em que os limites inferior e superior do envelope são dados respectivamente pelos vetores $t_{[i],[0,05m]}$ e $t_{[i],[0,95m]}$, para $i = 1, \ldots, n$, que denotam as réplicas correspondentes aos percentis 5 e 95 da m-ésima estatística de ordem. Entretanto, tal abordagem não será considerada no decorrer deste trabalho, uma vez que estudos preliminares indicaram que, se considerássemos como critério de rejeição do modelo a ocorrência de pelo menos um ponto fora, haveria uma quantidade considerável de pontos além das bandas do envelope mesmo sob o ajuste correto. Esse aspecto em particular não seria interessante para os estudos que conduziremos no Capítulo 5.

As Figuras 4.2 e 4.3 ilustram a inclusão do envelope simulado no gráfico normal de probabilidade segundo Atkinson, Everitt/Paula, Flack e Flores e Lee e Rhee para os mesmos dados gerados na Seção 4.1. Através das figuras, podemos observar alguns aspectos de cada um dos procedimentos. O envelope conforme Atkinson, presente na Figura 4.2 (esquerda), devido a um número reduzido de réplicas e a utilização de estatísticas de ordem extrema, apresenta bandas menos suaves que os demais métodos e amplitude das bandas menores que os métodos de Flack e Flores e Lee e Rhee. Para o procedimento de Everitt/Paula, Figura 4.2 (direita), em que são utilizadas mais réplicas, notamos que as bandas do envelope já são mais suaves e estáveis em relação a Atkinson. Por fim, a Figura 4.3 (esquerda) traz o envelope segundo Flack e Flores e a Figura 4.3 (direita) conforme Lee e Rhee, considerando para ambos $k \approx 2{,}06$. Podemos destacar que visualmente, ambos são relativamente parecidos, com bandas mais estáveis e suaves que os demais métodos, entretanto, a forma segundo Lee e Rhee demonstra ter uma amplitude ligeiramente maior, sendo inclusive, o envelope de maior amplitude dentre os avaliados.

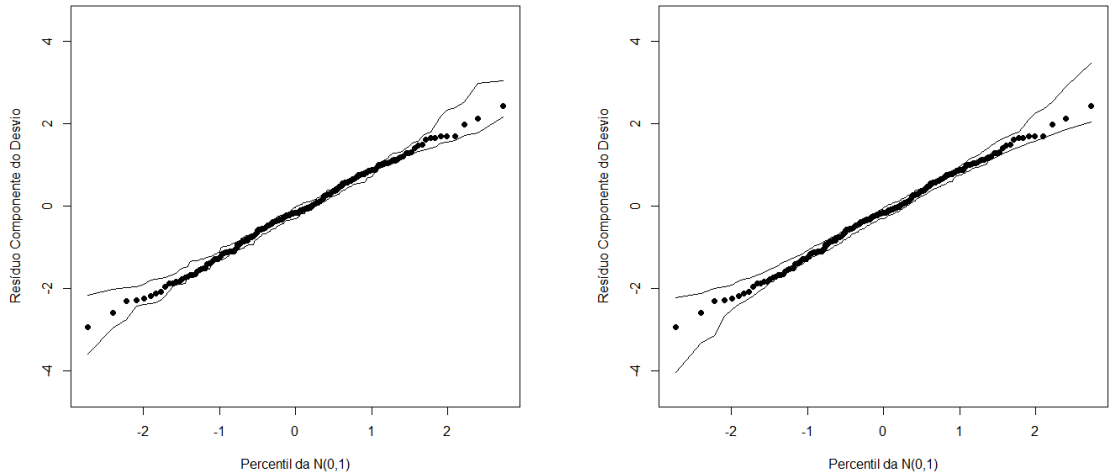

Figura 4.2: Envelopes simulados segundo Atkinson (esquerda) e Everitt/Paula (direita) para dados gerados assumindo um modelo de regressão gama.

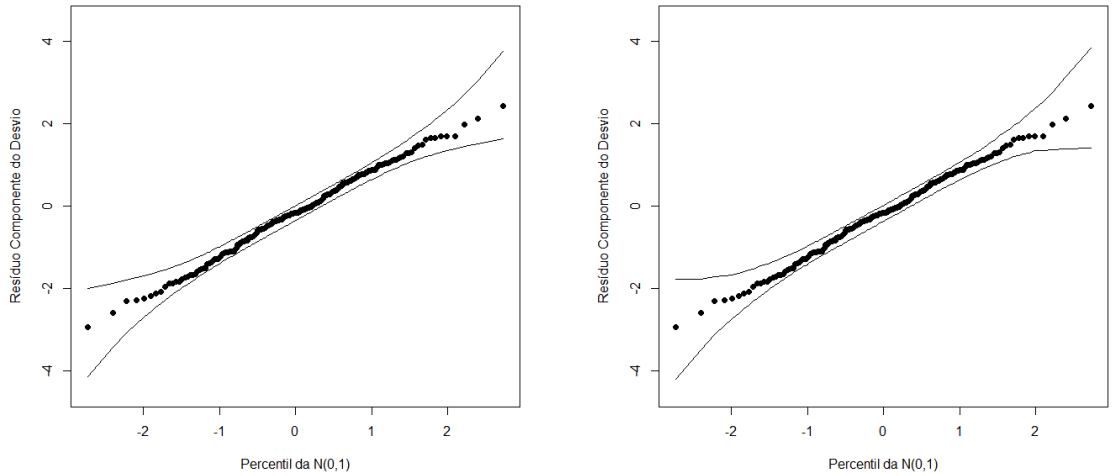

Figura 4.3: Envelopes simulados segundo Flack e Flores (esquerda) e Lee e Rhee (direita) para dados gerados assumindo um modelo de regressão gama, para $k \approx 2{,}06$.

4.2.5 Envelope simulado como teste formal

Uma outra abordagem para a construção do envelope consiste em encará-lo como um teste de hipóteses e, desta forma, determinar a grandeza de m, ou seja, o número de réplicas dos vetores de resíduos utilizados para a construção do envelope que forneça um nível de significância desejado. Duas propostas que fazem uso de tal raciocínio podem ser vistas em Ripley (1977) e Raubertas (1992).

4.3 Rejeição do modelo a partir do envelope

Conforme mencionado no Capítulo 1, a utilização do envelope simulado na verificação da distribuição dos dados em um modelo de regressão é bastante comum em trabalhos acadêmicos. Entretanto, embora o envelope forneça uma base para a flutuação dos resíduos, apenas isso não é o suficiente, já que a proporção de pontos fora do envelope para rejeição ou não de um modelo, por exemplo, é algo subjetivo.

4.3.1 Proposta de um método de rejeição do modelo a partir do envelope

Propomos aqui um método que auxilia na decisão de rejeitar ou não um modelo ajustado. A ideia para tal procedimento consiste em avaliar algumas quantidades características do envelope e desta forma, fornecer indícios sobre a plausabilidade da distribuição de probabilidade considerada no ajuste. O intuito sob esta óptica está em comparar a distribuição de um conjunto de dados observado com uma distribuição de referência especificada, a qual possui função distribuição denotada por G. Se F denota a verdadeira função distribuição da variável resposta, que é desconhecida, a questão de interesse é avaliar $H_0 : F(x) = G(x)$, para todo $x \in \mathbb{R}$.

A verificação da hipótese H_0 é feita através de duas quantidades obtidas a partir do envelope simulado, as quais são:

- A proporção de pontos fora do envelope;
- A distância entre a banda do envelope e o ponto mais afastado desse limite.
  - Esta medida simplesmente representa, quando há a ocorrência de resíduos fora do envelope, o valor absoluto da diferença entre o valor do resíduo e o valor da ordenada da banda mais próxima a ele. Se não houver pontos fora do envelope, essa distância é igual a zero.

A implementação do procedimento é feita através da construção do envelope segundo a proposta de Atkinson por meio dos seguintes passos:

1) Obtemos os limites do envelope conforme Atkinson para o modelo ajustado. Repetimos esse processo de obtenção 30 vezes, de tal forma que teremos 30 vetores de limites inferiores e 30 vetores de limites superiores.

2) O envelope apresentado ao usuário é dado pela média dos limites inferiores e superiores provenientes dos 30 passos replicados, no intuito de estabilizar as bandas do envelope.

3) Armazenamos do envelope do usuário a proporção de resíduos fora e a distância entre a banda do envelope e o ponto mais afastado desse limite, como ilustrado na Tabela 4.5.

4) Considerando o modelo ajustado, geramos a partir dele n observações $y_i^{(1)}$, para $i = 1, \ldots, n$, de tal forma que $\mathbf{y}^{(1)} = \{y_1^{(1)}, \ldots, y_n^{(1)}\}^\top$.

Tabela 4.5: Ilustração do armazenamento das quantidades características do envelope.

Prop. pontos fora	Maior dist. ponto envel.
a	b

5) Ajustamos $\mathbf{y}^{(1)}$ usando $\mathbf{X}$ e obtemos $\mathbf{t}^{(1)}$, com elementos $t_i^{(1)}$, para $i = 1, \ldots, n$. Referenciaremos o ajuste realizado neste passo como modelo suposto.

6) Geramos n observações $y_{i,1}^{(1)}$ a partir do modelo suposto no passo 5), de tal forma que teremos o vetor $\mathbf{y}_1^{(1)} = \{y_{1,1}^{(1)}, \ldots, y_{n,1}^{(1)}\}^\top$. Realizamos esse processo 19 vezes, no intuito de criar para os resíduos do modelo suposto um envelope segundo a proposta de Atkinson e, assim obtemos $\mathbf{y}_1^{(1)} = \{y_{1,1}^{(1)}, \ldots, y_{n,1}^{(1)}\}^\top, \ldots, \mathbf{y}_{19}^{(1)} = \{y_{1,19}^{(1)}, \ldots, y_{n,19}^{(1)}\}^\top$, conforme Tabela 4.6.

Tabela 4.6: Réplicas provenientes do modelo suposto.

Réplica $\mathbf{y}_1^{(1)}$	Réplica $\mathbf{y}_2^{(1)}$	...	Réplica $\mathbf{y}_{18}^{(1)}$	Réplica $\mathbf{y}_{19}^{(1)}$
$y_{1,1}^{(1)}$	$y_{2,2}^{(1)}$	$\cdots$	$y_{1,18}^{(1)}$	$y_{1,19}^{(1)}$
$y_{2,1}^{(1)}$	$y_{2,2}^{(1)}$	$\cdots$	$y_{2,18}^{(1)}$	$y_{2,19}^{(1)}$
$y_{3,1}^{(1)}$	$y_{3,2}^{(1)}$	$\cdots$	$y_{3,18}^{(1)}$	$y_{3,19}^{(1)}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$y_{n-2,1}^{(1)}$	$y_{n-2,2}^{(1)}$	$\cdots$	$y_{n-2,18}^{(1)}$	$y_{n-2,19}^{(1)}$
$y_{n-1,1}^{(1)}$	$y_{n-1,2}^{(1)}$	$\cdots$	$y_{n-1,18}^{(1)}$	$y_{n-1,19}^{(1)}$
$y_{n,1}^{(1)}$	$y_{n,2}^{(1)}$	$\cdots$	$y_{n,18}^{(1)}$	$y_{n,19}^{(1)}$

7) Para cada $\mathbf{y}_1^{(1)}, \mathbf{y}_2^{(1)}, \ldots, \mathbf{y}_{18}^{(1)}, \mathbf{y}_{19}^{(1)}$, proveniente do modelo suposto, realizamos o ajuste usando $\mathbf{X}$ e obtemos os respectivos vetores de resíduos $\mathbf{t}_1^{(1)}, \mathbf{t}_2^{(1)}, \ldots, \mathbf{t}_{18}^{(1)}, \mathbf{t}_{19}^{(1)}$. Através do procedimento de Atkinson, construímos o envelope para $t_{[i]}^{(1)}$, com $i = 1, \ldots, n$, como verificamos na Tabela 4.7.

Tabela 4.7: Obtendo os limites do envelope para o modelo suposto.

	⇓	⇓		⇓	⇓			
	Resíduo $\mathbf{t}_1^{(1)}$	Resíduo $\mathbf{t}_2^{(1)}$	...	Resíduo $\mathbf{t}_{18}^{(1)}$	Resíduo $\mathbf{t}_{19}^{(1)}$	Mínimo	Resíduo $\mathbf{t}^{(1)}$	Máximo
$\Rightarrow$	$t_{[1],1}^{(1)}$	$t_{[1],2}^{(1)}$	$\cdots$	$t_{[1],18}^{(1)}$	$t_{[1],19}^{(1)}$	$t_{[1],[1]}^{(1)}$	$t_{[1]}^{(1)}$	$t_{[1],[19]}^{(1)}$
$\Rightarrow$	$t_{[2],1}^{(1)}$	$t_{[2],2}^{(1)}$	$\cdots$	$t_{[2],18}^{(1)}$	$t_{[2],19}^{(1)}$	$t_{[2],[1]}^{(1)}$	$t_{[2]}^{(1)}$	$t_{[2],[19]}^{(1)}$
$\Rightarrow$	$t_{[3],1}^{(1)}$	$t_{[3],2}^{(1)}$	$\cdots$	$t_{[3],18}^{(1)}$	$t_{[3],19}^{(1)}$	$t_{[3],[1]}^{(1)}$	$t_{[3]}^{(1)}$	$t_{[3],[19]}^{(1)}$
	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\Rightarrow$	$t_{[n-2],1}^{(1)}$	$t_{[n-2],2}^{(1)}$	$\cdots$	$t_{[n-2],18}^{(1)}$	$t_{[n-2],19}^{(1)}$	$t_{[n-2],[1]}^{(1)}$	$t_{[n-2]}^{(1)}$	$t_{[n-2],[19]}^{(1)}$
$\Rightarrow$	$t_{[n-1],1}^{(1)}$	$t_{[n-1],2}^{(1)}$	$\cdots$	$t_{[n-1],18}^{(1)}$	$t_{[n-1],19}^{(1)}$	$t_{[n-1],[1]}^{(1)}$	$t_{[n-1]}^{(1)}$	$t_{[n-1],[19]}^{(1)}$
$\Rightarrow$	$t_{[n],1}^{(1)}$	$t_{[n],2}^{(1)}$	$\cdots$	$t_{[n],18}^{(1)}$	$t_{[n],19}^{(1)}$	$t_{[n],[1]}^{(1)}$	$t_{[n]}^{(1)}$	$t_{[n],[19]}^{(1)}$

8) Obtemos do envelope criado para o modelo suposto a proporção de resíduos fora do envelope e a distância entre a banda do envelope e o ponto mais afastado desse limite, como apresentado na Tabela 4.8.

Tabela 4.8: Armazenamento das quantidades características do envelope para o modelo suposto.

Prop. pontos fora	Maior dist. ponto envel.
$a^{(1)}$	$b^{(1)}$

9) Repetimos r vezes os passos 4) a 8), de tal forma que teremos r modelos supostos e armazenaremos para cada um deles as duas quantidades características obtidas a partir dos seus respectivos envelopes.

10) Armazenamos a média das 2 quantidades fornecidas nas r repetições. Posteriormente, acrescentamos um fator multiplicativo a elas, conforme Tabela 4.9 (a média associada ao fator multiplicativo denominamos de quantidade base). Comparamos as quantidades bases com as obtidas no passo 3). O critério de rejeição é dado da seguinte forma: se, em pelo menos uma das medidas, os valores no passo 3) forem maiores que as quantidades bases, o modelo deve ser rejeitado.

Tabela 4.9: Ilustração da obtenção das quantidades bases.

	Prop. pontos fora	Maior dist. ponto envel.
	$a^{(1)}$	$b^{(1)}$
	$a^{(2)}$	$b^{(2)}$
	$\vdots$	$\vdots$
	$a^{(r-1)}$	$b^{(r-1)}$
	$a^{(r)}$	$b^{(r)}$
Base	2,5*média($a^{(.)}$)	3,5*média($b^{(.)}$)

Para a avaliação do procedimento, faz-se necessário definir a taxa de rejeição nula do método. Tal quantidade é análoga ao erro tipo I em um teste de hipótese, porém, como estamos obtendo esse valor via simulação, atribuímos outra nomenclatura. A taxa de rejeição nula nada mais é que a proporção de vezes nas réplicas de Monte Carlo que o método rejeita a distribuição G, dado que ela realmente é a distribuição que gerou os dados. Por meio de diversos estudos preliminares, verificamos que a utilização de r = 50 fornece uma taxa de rejeição nula estável por volta de 0,05, conforme poderá ser visto no Capítulo 5.

Já a taxa de rejeição não nula é uma quantidade análoga ao poder do teste na conjuntura dos testes de hipótese, que nada mais é que a proporção de vezes que o procedimento rejeita a distribuição G dado que ela não é a distribuição que gerou os dados.

Note que no passo 10 acrescentamos um fator multiplicativo para a média da proporção de pontos fora e para a distância entre a banda do envelope e o ponto mais afastado desse limite. Esse artifício foi realizado, pois verificamos através de simulações que,

em média, embora as quantidades obtidas do envelope do usuário e as armazenadas do envelope do modelo suposto sejam praticamente iguais, os fatores multiplicativos para cada uma das medidas foram escolhidos de forma a garantir a estabilidade da taxa de rejeição nula em torno de 0,05. Uma maneira mais simples para determinar as quantidades bases seria utilizar o máximo das r repetições para cada uma das medidas armazenadas. Entretanto, por meio de simulações, verificou-se em tal abordagem que a decisão do método de rejeição se torna instável em um mesmo conjunto de dados, ou seja, o resultado é dependente da aleatorização utilizada. Esse aspecto se verifica pois o máximo é uma estatística de ordem extrema e portanto, suscetível ao aparecimento de valores consideravelmente diferentes.

Como o procedimento proposto é essencialmente baseado na forma de construção segundo Atkinson, as bandas do envelope são menos estáveis e suaves que dos demais métodos. Desta forma, repetir o processo de obtenção dos limites do envelope e fornecer ao usuário um envelope final dado pela média dessas bandas, assim como replicar a construção do envelope para o modelo suposto diversas vezes, foi o subterfúgio utilizado para tornar mais estável tanto os limites do envelope do usuário quanto as quantidades bases e, consequentemente, também a decisão do método em um mesmo banco de dados. Vale ressaltar que, tanto os valores utilizados para as réplicas do envelope, quanto as quantidades atribuídas aos fatores multiplicativos, são particulares para este método que propomos. Caso sejam feitas alterações, novos valores deverão ser investigados considerando-se estimativas das taxas de rejeição obtidas através de estudos de simulação.

Além da proporção de pontos fora do envelope e da distância entre a banda do envelope e o ponto mais afastado desse limite, outras medidas poderiam ser consideradas para compor o método de rejeição do modelo. Dentre elas, podemos citar a quantidade máxima de pontos consecutivos além das bandas do envelope e a quantidade máxima de pontos consecutivos que saem nas extremidades inferior ou superior do envelope. Ambas as medidas foram testadas, mas verificamos que elas tornavam o procedimento instável para uma mesma amostra, uma vez que pequenas oscilações na obtenção das bandas causavam discrepâncias consideráveis nas medidas. Desta forma, optamos em não utilizá-las na construção do método.

4.3.2 Outras abordagens para a rejeição do modelo a partir do envelope

No intuito de comparar o procedimento proposto com demais métodos já existentes, um possível critério para a rejeição do modelo baseado em um envelope é rejeitar o mesmo se um ponto estiver fora dos limites do envelope. Utilizando essa ideia, Flack and Flores (1989) mostraram que seu método de construção do envelope, quando o ajuste é o correto, tem menor probabilidade de rejeitar o modelo do que o envelope de Atkinson. Sendo assim, consideraremos tal critério para a rejeição do modelo para os procedimentos de Flack e Flores e Lee e Rhee. Discutiremos mais a respeito de tais comparações no Capítulo

5. Os métodos citados na Subseção 4.2.5 não serão considerados no estudo, dado que é uma tarefa relativamente complexa definir a grandeza de m, o número de réplicas dos vetores de resíduos utilizados para a construção do envelope, que forneça um nível de significância desejado em um estudo de simulação.

Os métodos de rejeição baseados em único ponto fora, construídos a partir das propostas de Flack e Flores e Lee e Rhee, necessitam, respectivamente, de 19000 e 2000 ajustes para a sua obtenção. Já o método proposto se destaca positivamente nesse quesito, uma vez que os 1570 modelos gerados para a sua criação representam um custo computacional muito menor que os demais.

Capítulo 5

Estudo de simulação

Neste capítulo, discutimos os estudos de simulação via Monte Carlo que foram conduzidos considerando os modelos presentes nos Capítulos 2 e 3. Tais estudos foram realizados utilizando-se o *software* R (R Core Team, 2018) através dos pacotes *gamlss* e *statmod*. Na Seção 5.1 apresentamos os resultados obtidos com os modelos lineares generalizados para variáveis com distribuição contínua e discreta; ao passo que na Seção 5.2 abordamos os resultados referentes aos modelos de regressão para taxas e proporções. Em ambas as classes presentes no estudo, comparamos a performance dos métodos de rejeição baseados na construção do envelope simulado segundo Flack e Flores, Lee e Rhee, além do procedimento que propomos na Seção 4.3. Outra comparação que apresentamos é referente ao uso do gráfico normal e meio normal com envelope simulado conforme as propostas de Atkinson e Everitt/Paula. Para todos os casos estudados, consideramos três tamanhos amostrais, sendo n = 50, 100 e 200. A seguir discutiremos com mais detalhes a avaliação numérica referente aos MLG.

5.1 Avaliação numérica: Modelos lineares generalizados

No que concerne aos modelos lineares generalizados, utilizamos no estudo das distribuições contínuas a gama e a normal inversa. Para cada uma delas definimos seis cenários, sendo que nos cinco primeiros consideramos a função de ligação logarítmica e no último, a função de ligação raiz quadrada. Além disso, utilizamos em todos os cenários duas variáveis preditoras. No primeiro cenário e nos três últimos, as covariáveis foram geradas independentemente de uniformes padrão. Já para o cenário II, geramos x_{i1} de uma distribuição normal ($x_{i1} \sim N(0{,}5; 0{,}25^2)$) e x_{i2} de uma distribuição normal inversa ($x_{i2} \sim NI(0{,}4; 2)$), enquanto que para o cenário III, mantivemos x_{i1} e geramos x_{i2} de uma distribuição gama ($x_{i2} \sim GA(0{,}4; 1)$). É importante ressaltar que as covariáveis foram geradas e posteriormente mantidas fixas em todas as réplicas de Monte Carlo realizadas.

Por fim, consideramos diferentes valores para o parâmetro de dispersão ϕ. Nos três primeiros cenários, utilizamos $\phi = 0{,}01$ se a distribuição fosse normal inversa e $\phi = 0{,}25$ se a distribuição fosse gama. Já nos cenários IV e V diminuímos e aumentamos tais valores,

respectivamente, a fim de avaliar o impacto dessas alterações. Considerando pequenas modificações, os cenários I a V são aqueles utilizados por Scudilio and Pereira (2017).

Para facilitar a visualização dos cenários, apresentamos nas Tabelas 5.1 e 5.2 a descrição dos mesmos para os modelos de regressão normal inversa e gama, respectivamente. No primeiro cenário (cenários I-a e I-b), atribuímos ao vetor de parâmetros os valores $\boldsymbol{\beta} = (3;2;1)$ e, consequentemente, $\mu \in (20{,}086;403{,}429)$. Outras informações relevantes podem ser obtidas a partir das tabelas.

Tabela 5.1: Descrição dos cenários para o modelo de regressão normal inversa.

Cenário	Função de ligação	Coeficientes	Covariáveis	μ	ϕ
I-a	$log(\mu_i)$	$\beta_0 = 3, \beta_1 = 2, \beta_2 = 1$	$x_{i1} \sim U(0,1)$ e $x_{i2} \sim U(0,1)$	$(20{,}086; 403{,}429)$	$0,01$
II-a	$log(\mu_i)$	$\beta_0 = 3, \beta_1 = 2, \beta_2 = 1$	$x_{i1} \sim N(0{,}5; 0{,}25^2)$ e $x_{i2} \sim NI(0{,}4; 2)$	$(13{,}728; 986{,}463)$	$0,01$
III-a	$log(\mu_i)$	$\beta_0 = 3, \beta_1 = 2, \beta_2 = 1$	$x_{i1} \sim N(0{,}5; 0{,}25^2)$ e $x_{i2} \sim GA(0{,}4; 1)$	$(12{,}463; 827{,}028)$	$0,01$
IV-a	$log(\mu_i)$	$\beta_0 = 3, \beta_1 = 2, \beta_2 = 1$	$x_{i1} \sim U(0,1)$ e $x_{i2} \sim U(0,1)$	$(20{,}086; 403{,}429)$	$0,005$
V-a	$log(\mu_i)$	$\beta_0 = 3, \beta_1 = 2, \beta_2 = 1$	$x_{i1} \sim U(0,1)$ e $x_{i2} \sim U(0,1)$	$(20{,}086; 403{,}429)$	$0,02$
VI-a	$\sqrt{\mu_i}$	$\beta_0 = 4, \beta_1 = 8, \beta_2 = 6$	$x_{i1} \sim U(0,1)$ e $x_{i2} \sim U(0,1)$	$(16{,}000; 324{,}000)$	$0,01$

Tabela 5.2: Descrição dos cenários para o modelo de regressão gama.

Cenário	Função de ligação	Coeficientes	Covariáveis	μ	ϕ
I-b	$log(\mu_i)$	$\beta_0 = 3, \beta_1 = 2, \beta_2 = 1$	$x_{i1} \sim U(0,1)$ e $x_{i2} \sim U(0,1)$	$(20{,}086; 403{,}429)$	$0,25$
II-b	$log(\mu_i)$	$\beta_0 = 3, \beta_1 = 2, \beta_2 = 1$	$x_{i1} \sim N(0{,}5; 0{,}25^2)$ e $x_{i2} \sim NI(0{,}4; 2)$	$(13{,}728; 986{,}463)$	$0,25$
III-b	$log(\mu_i)$	$\beta_0 = 3, \beta_1 = 2, \beta_2 = 1$	$x_{i1} \sim N(0{,}5; 0{,}25^2)$ e $x_{i2} \sim GA(0{,}4; 1)$	$(12{,}463; 827{,}028)$	$0,25$
IV-b	$log(\mu_i)$	$\beta_0 = 3, \beta_1 = 2, \beta_2 = 1$	$x_{i1} \sim U(0,1)$ e $x_{i2} \sim U(0,1)$	$(20{,}086; 403{,}429)$	$0,125$
V-b	$log(\mu_i)$	$\beta_0 = 3, \beta_1 = 2, \beta_2 = 1$	$x_{i1} \sim U(0,1)$ e $x_{i2} \sim U(0,1)$	$(20{,}086; 403{,}429)$	$0,5$
VI-b	$\sqrt{\mu_i}$	$\beta_0 = 4, \beta_1 = 8, \beta_2 = 6$	$x_{i1} \sim U(0,1)$ e $x_{i2} \sim U(0,1)$	$(16{,}000; 324{,}000)$	$0,25$

Já para caso discreto, utilizamos as distribuições poisson e binomial negativa, em que foram definidos três cenários. Para os dois primeiros, atribuímos ao preditor a função de ligação logarítmica e para o último, a função de ligação raiz quadrada. No que diz respeito as covariáveis, nos cenários I e III geramos as mesmas de duas uniformes padrão independentes. Para o cenário II, utilizamos para x_{i1} uma distribuição normal ($x_{i1} \sim N(0{,}5; 0{,}25^2)$) e para x_{i2} uma distribuição exponencial ($x_{i2} \sim Exp(0{,}15)$). Vale destacar mais uma vez que as variáveis preditoras foram geradas e, posteriormente, mantidas fixas em todas as réplicas de Monte Carlo. Além disso, consideramos $\sigma = 0{,}5$ para a geração da variável resposta com distribuição binomial negativa.

A Tabela 5.3 traz detalhes adicionais referentes a simulação. No cenário I, por exemplo, atribuímos ao vetor de parâmetros os valores $\boldsymbol{\beta} = (0{,}5;2;1)$ e consequentemente, $\mu \in (1{,}649; 33{,}115)$.

Tabela 5.3: Descrição dos cenários para o modelo de regressão poisson e binomial negativa.

Cenário	Função de ligação	Coeficientes	Covariáveis	μ
I	$log(\mu_i)$	$\beta_0 = 0{,}5, \beta_1 = 2, \beta_2 = 1$	$x_{i1} \sim U(0,1)$ e $x_{i2} \sim U(0,1)$	$(1{,}649; 33{,}115)$
II	$log(\mu_i)$	$\beta_0 = 0{,}5, \beta_1 = 2, \beta_2 = 1$	$x_{i1} \sim N(0{,}5; 0{,}25^2)$ e $x_{i2} \sim Exp(0{,}15)$	$(1{,}109; 29{,}256)$
III	$\sqrt{(\mu_i)}$	$\beta_0 = 1, \beta_1 = 2, \beta_2 = 3$	$x_{i1} \sim U(0,1)$ e $x_{i2} \sim U(0,1)$	$(1{,}000; 36{,}000)$

5.1.1 Comparação entre os procedimentos de rejeição do modelo

No intuito de comparar o procedimento proposto com outros métodos de rejeição do modelo, consideramos para as formas de construção do envelope segundo Flack e Flores e Lee e Rhee o critério que rejeitará o modelo se um ponto estiver fora do envelope. Desta maneira, avaliamos no estudo quatro situações possíveis para os MLG com variável resposta contínua e duas para os MLG com variável resposta discreta, as quais respectivamente são:

- $NI \rightarrow NI$ representa a ocasião em que geramos a amostra de uma distribuição normal inversa e construímos o envelope considerando uma distribuição normal inversa;
- $NI \rightarrow GA$ representa a ocasião em que geramos a amostra de uma distribuição normal inversa e construímos o envelope considerando uma distribuição gama;
- $GA \rightarrow GA$ representa a ocasião em que geramos a amostra de uma distribuição gama e construímos o envelope também considerando uma distribuição gama;
- $GA \rightarrow NI$ representa a ocasião em que geramos a amostra de uma distribuição gama e construímos o envelope considerando uma distribuição normal inversa;
- $PO \rightarrow PO$ representa a ocasião em que geramos a amostra de uma distribuição poisson e construímos o envelope considerando uma distribuição poisson;
- $BN \rightarrow PO$ representa a ocasião em que geramos a amostra de uma distribuição binomial negativa e construímos o envelope considerando uma distribuição poisson.

Para o procedimento proposto e do Lee e Rhee, realizamos 2000 réplicas via Monte Carlo do processo de rejeição do modelo. Já para o método baseado na construção de Flack e Flores, devido ao tempo computacional mais elevado, efetuamos 500 repetições via Monte Carlo. Para a simulação, fizemos uso de $k \approx 2{,}06$, dado que por meio de estudos preliminares, avaliamos que o mesmo apresentou melhores resultados que $k = 1{,}5$ em termos da taxa de rejeição nula. Além disso, utilizamos o resíduo componente do desvio para a obtenção das taxas de rejeição nula e não nula nos três procedimentos.

As Tabelas 5.4 a 5.9 apresentam os resultados para as distribuições gama e normal inversa contendo as taxas de rejeição nula e rejeição não nula para os procedimentos de Flack e Flores, Lee e Rhee e o proposto, referente aos cenários de I a VI. De maneira geral, notamos que nas situações $NI \rightarrow NI$ e $GA \rightarrow GA$, a performance dos procedimentos de Flack e Flores e de Lee e Rhee parecem ser influenciados pelo tamanho da amostra, uma vez que a taxa de rejeição nula tende a ser maior conforme aumenta-se a quantidade de observações. Em contrapartida, o método proposto aparenta ter uma taxa de rejeição nula relativamente estável, que oscila mais ou menos em torno de 0,05 independente do tamanho da amostra. No que se refere a taxa de rejeição não nula, observada através das condições $NI \rightarrow GA$ e $GA \rightarrow NI$, verificamos que, como esperado em todos os cenários,

os procedimentos tendem a possuir taxa de rejeição não nula maior conforme aumenta-se a quantidade de observações. Isso significa que quanto maior o tamanho amostral, mais fácil torna-se para os métodos indicarem quando a distribuição incorreta é ajustada.

Os casos em que um dos métodos possui melhor performance (taxa de rejeição nula menor e simultaneamente maior taxa de rejeição não nula) que pelo menos um dos demais estão evidenciados em negrito. Observamos que o método proposto é o único que possui casos evidenciados em negrito e isso ocorre para pelo menos dois tamanhos amostrais, considerando todos os cenários avaliados. Dessa forma, o método proposto parece apresentar melhor performance que os demais em MLG com variável resposta contínua. Podemos destacar que, para tamanhos amostrais pequenos, considerando n = 50, o procedimento de Lee e Rhee é o que apresenta as menores taxas de rejeição nula dentre os três métodos vistos. Entretanto, em geral, também é aquele com menor taxa de rejeição não nula em amostras pequenas. Já o procedimento de Flack e Flores, embora seja o método que apresente taxas de rejeição não nula mais próximas ao método proposto, em geral é aquele em que a taxa de rejeição nula mais cresce com o tamanho da amostra.

A mudança da distribuição das variáveis preditoras (cenários II e III), do parâmetro de dispersão (cenários IV e V) e da função de ligação (cenário VI) não parecem afetar substancialmente a taxa de rejeição nula nos três métodos. Já para a taxa de rejeição não nula, notamos que ela se altera mais de um cenário para o outro. Entretanto, tal comportamento pode ser resultado das distribuições normal inversa e gama serem mais semelhantes entre si em alguns cenários do que em outros.

Tabela 5.4: Resultados de simulação para os cenários I-a e I-b - Taxas de rejeição nula e rejeição não nula.

n	Método	$NI \to NI$	$NI \to GA$	$GA \to GA$	$GA \to NI$
50	Flack e Flores	0,062	0,384	0,046	0,414
	Lee e Rhee	0,025	0,296	0,028	0,377
	Proposto	**0,051**	**0,430**	**0,036**	**0,459**
100	Flack e Flores	0,078	0,728	0,064	0,768
	Lee e Rhee	0,041	0,610	0,050	0,711
	Proposto	0,046	0,718	**0,037**	**0,776**
200	Flack e Flores	0,080	0,950	0,102	0,968
	Lee e Rhee	0,060	0,903	0,065	0,955
	Proposto	**0,050**	**0,952**	**0,051**	**0,964**

Tabela 5.5: Resultados de simulação para os cenários II-a e II-b - Taxas de rejeição nula e rejeição não nula.

n	Método	$NI \to NI$	$NI \to GA$	$GA \to GA$	$GA \to NI$
50	Flack e Flores	0,052	0,318	0,048	0,388
	Lee e Rhee	0,024	0,250	0,028	0,357
	Proposto	0,056	0,359	**0,047**	**0,414**
100	Flack e Flores	0,064	0,674	0,064	0,816
	Lee e Rhee	0,040	0,547	0,046	0,700
	Proposto	0,052	0,608	0,054	0,782
200	Flack e Flores	0,088	0,900	0,098	0,976
	Lee e Rhee	0,051	0,861	0,066	0,922
	Proposto	**0,067**	**0,919**	**0,040**	**0,953**

Tabela 5.6: Resultados de simulação para os cenários III-a e III-b - Taxas de rejeição nula e rejeição não nula.

n	Método	$NI \to NI$	$NI \to GA$	$GA \to GA$	$GA \to NI$
50	Flack e Flores	0,048	0,314	0,060	0,486
	Lee e Rhee	0,025	0,254	0,030	0,356
	Proposto	0,060	0,356	0,045	0,468
100	Flack e Flores	0,070	0,630	0,094	0,790
	Lee e Rhee	0,035	0,550	0,039	0,739
	Proposto	0,065	0,607	**0,052**	**0,792**
200	Flack e Flores	0,080	0,940	0,104	0,974
	Lee e Rhee	0,051	0,897	0,061	0,962
	Proposto	0,055	0,919	**0,061**	**0,978**

Tabela 5.7: Resultados de simulação para os cenários IV-a e IV-b - Taxas de rejeição nula e rejeição não nula.

n	Método	$NI \to NI$	$NI \to GA$	$GA \to GA$	$GA \to NI$
50	Flack e Flores	0,068	0,258	0,056	0,254
	Lee e Rhee	0,025	0,197	0,027	0,250
	Proposto	0,053	0,258	**0,035**	**0,298**
100	Flack e Flores	0,090	0,568	0,066	0,614
	Lee e Rhee	0,040	0,431	0,053	0,522
	Proposto	0,042	0,470	**0,040**	**0,579**
200	Flack e Flores	0,072	0,840	0,094	0,890
	Lee e Rhee	0,055	0,760	0,065	0,839
	Proposto	**0,047**	**0,778**	**0,051**	**0,856**

Tabela 5.8: Resultados de simulação para os cenários V-a e V-b - Taxas de rejeição nula e rejeição não nula.

n	Método	$NI \to NI$	$NI \to GA$	$GA \to GA$	$GA \to NI$
50	Flack e Flores	0,060	0,530	0,038	0,640
	Lee e Rhee	0,022	0,428	0,026	0,608
	Proposto	**0,057**	**0,671**	0,040	0,693
100	Flack e Flores	0,074	0,896	0,060	0,924
	Lee e Rhee	0,032	0,812	0,048	0,922
	Proposto	**0,049**	**0,922**	**0,041**	**0,951**
200	Flack e Flores	0,072	0,998	0,092	1,000
	Lee e Rhee	0,049	0,988	0,059	0,998
	Proposto	**0,053**	**0,999**	**0,052**	**0,999**

Tabela 5.9: Resultados de simulação para os cenários VI-a e VI-b - Taxas de rejeição nula e rejeição não nula.

n	Método	$NI \to NI$	$NI \to GA$	$GA \to GA$	$GA \to NI$
50	Flack e Flores	0,048	0,458	0,038	0,392
	Lee e Rhee	0,028	0,338	0,030	0,357
	Proposto	0,061	0,537	0,046	0,437
100	Flack e Flores	0,056	0,796	0,052	0,730
	Lee e Rhee	0,040	0,682	0,051	0,678
	Proposto	**0,046**	**0,832**	**0,041**	**0,736**
200	Flack e Flores	0,068	0,976	0,094	0,956
	Lee e Rhee	0,057	0,950	0,064	0,946
	Proposto	**0,048**	**0,988**	**0,047**	**0,952**

Os resultados contendo as taxas de rejeição nula e não nula para os três cenários simulados, avaliando as distribuições poisson e binomial negativa, estão presentes nas Tabelas 5.10 a 5.12. É possível notar em todos os cenários que a taxa de rejeição nula nos procedimentos de Flack e Flores e Lee e Rhee demonstra, assim como nos MLG para distribuições contínuas, ser influenciada pelo tamanho amostral. Já no procedimento proposto, verificamos que a taxa de rejeição nula flutua por volta de 0,07 e 0,08, para tamanhos de amostra menores, e diminui conforme aumenta-se o número de observações, com oscilação em torno de 0,06. Para a taxa de rejeição não nula, observada através da condição $BN \rightarrow PO$, nos três cenários os procedimentos tendem a possuir taxa de rejeição não nula maior conforme aumenta-se a quantidade de observações. As modificações feitas quanto a distribuição das variáveis preditoras (cenário II) e a função de ligação (cenário III) também demonstram influenciar apenas a taxa de rejeição não nula dos procedimentos. É importante ressaltar que a performance do método proposto nos MLG com variável resposta discreta não parece ser superior aos dos demais métodos como ocorre nos MLG com variável resposta contínua. Uma possível explicação para isso se deve ao fato dos resíduos, em particular o resíduo componente do desvio, não possuírem distribuição contínua para o caso dos MLG com variável resposta discreta (Dunn and Smyth, 1996).

Tabela 5.10: Resultados de simulação para o cenário I - Taxas de rejeição nula e rejeição não nula.

n	Método	$PO \rightarrow PO$	$BN \rightarrow PO$
50	Flack e Flores	**0,064**	**0,506**
	Lee e Rhee	0,042	0,460
	Proposto	0,069	0,477
100	Flack e Flores	**0,068**	**0,784**
	Lee e Rhee	0,049	0,695
	Proposto	0,071	0,716
200	Flack e Flores	0,090	0,956
	Lee e Rhee	0,053	0,934
	Proposto	0,058	0,938

Tabela 5.11: Resultados de simulação para o cenário II - Taxas de rejeição nula e rejeição não nula.

n	Método	$PO \rightarrow PO$	$BN \rightarrow PO$
50	Flack e Flores	0,080	0,530
	Lee e Rhee	0,033	0,445
	Proposto	0,079	0,502
100	Flack e Flores	0,076	0,770
	Lee e Rhee	0,058	0,705
	Proposto	0,069	0,705
200	Flack e Flores	0,080	0,954
	Lee e Rhee	0,049	0,923
	Proposto	0,063	0,926

Tabela 5.12: Resultados de simulação para o cenário III - Taxas de rejeição nula e rejeição não nula.

n	Método	$PO \rightarrow PO$	$BN \rightarrow PO$
50	Flack e Flores	**0,054**	**0,504**
	Lee e Rhee	0,040	0,440
	Proposto	0,072	0,492
100	Flack e Flores	**0,066**	**0,776**
	Lee e Rhee	0,049	0,702
	Proposto	0,069	0,707
200	Flack e Flores	0,092	0,952
	Lee e Rhee	0,058	0,934
	Proposto	0,063	0,939

5.1.2 Comparação entre o gráfico normal e meio normal probabilístico com envelope simulado

Após a comparação dos métodos de rejeição do modelo, verificamos agora a performance dos gráficos normal e meio normal probabilístico com envelope simulado através dos cenários de I a VI. Para isso, utilizamos 2000 réplicas via Monte Carlo do processo de geração do envelope conforme os procedimentos de construção de Atkinson e Everitt/Paula, em que para cada uma dessas réplicas, computamos estatísticas referentes a proporção de pontos fora do envelope, a distância entre a banda do envelope e o ponto mais afastado desse limite e a proporção máxima de pontos consecutivos além das bandas do envelope, que nada mais é que a divisão da quantidade máxima de pontos consecutivos fora do envelope pelo tamanho amostral. Vale ressaltar que neste estudo também fizemos uso do resíduo componente do desvio.

Os resultados contendo as estatísticas descritas estão presentes nas Tabelas 5.13 e 5.14 para o cenário I e nas Tabelas de A.1 a A.10 para os demais cenários (apresentadas no Apêndice A). Além disso, incluímos aos resultados a taxa de rejeição nula e não nula do procedimento proposto, no intuito de fornecer uma base comparativa às formas tradicionais de obtenção do envelope. Outra métrica que acrescentamos foi a proporção de vezes em que pelo menos um ponto estava fora do envelope nas 2000 réplicas, segundo Atkinson e Everitt/Paula. Temos que, em todos os cenários, as médias das medidas avaliadas sob o modelo correto ($NI \rightarrow NI$ e $GA \rightarrow GA$) demonstram oscilar ao redor de um determinado patamar, que independe do tamanho amostral, da distribuição gerada (normal inversa ou gama) ou do gráfico de probabilidade utilizado. No que diz respeito a média da proporção de pontos fora do envelope, observamos que a mesma varia em torno de 0,09 e 0,1 para a forma de obtenção de Atkinson e por volta de 0,04 e 0,05 para o envelope segundo Everitt/Paula. É interessante notar que tais verificações são similares as especificações que justificam a forma de construção dos métodos, conforme argumentado na Subseção 4.2.1.

Ainda sob o modelo correto, outro ponto de destaque diz respeito a proporção máxima de pontos consecutivos fora do envelope. É possível notar que independente do gráfico de probabilidade utilizado, a média desta quantidade oscila por volta de 0,05 e 0,06 para o método de Atkinson e em torno de 0,02 e 0,03 para o procedimento segundo Everitt/Paula. Já sob o modelo incorreto, observamos que a média de todas as três quantidades utilizadas tendem, como o esperado, a crescerem quando se aumenta o tamanho da amostra.

Conduzindo a análise pela distribuição gerada, iniciamos pelo caso em que a variável resposta possui distribuição normal inversa. Em geral, verificamos que o gráfico normal probabilístico com envelope simulado demonstra ter melhor performance que o gráfico meio normal de probabilidade, tanto para o método segundo Atkinson quanto conforme Everitt/Paula, se levarmos em consideração as médias da proporção de pontos fora e da proporção máxima de pontos consecutivos além das bandas do envelope. Quando avaliamos o ajuste correto, vemos que ambas as medidas de proporção são próximas para os gráficos de probabilidade normal e meio normal, entretanto, para o gráfico normal a média tende a ser menor. Considerando agora o modelo incorreto, ou seja, $NI \rightarrow GA$, observamos que o gráfico normal apresenta um desempenho substancialmente melhor que o gráfico meio normal, com base na média da proporção de pontos fora e na média da proporção máxima de pontos consecutivos além das bandas do envelope.

Quando geramos a variável resposta de uma distribuição gama, não é possível verificar nenhuma tendência muito clara sobre qual gráfico de probabilidade apresenta melhor desempenho ao longo dos cenários. Isso ocorre pois sob o modelo correto, o gráfico normal apresenta média da proporção de pontos fora e da proporção de pontos consecutivos fora menor que o meio normal, segundo Atkinson e Everitt/Paula, enquanto o meio normal possui média das mesmas medidas maior para o ajuste incorreto. Já para a média da distância entre a banda do envelope e o ponto mais afastado desse limite, esse comportamento entre os gráficos de probabilidade é observado de maneira contrária, ou seja, menor média sob o ajuste correto para o gráfico meio normal e sob o modelo incorreto, maior média da distância para o gráfico meio normal.

No que se refere as formas de construção do envelope, o procedimento segundo Atkinson é o mais rápido computacionalmente e, sob o ajuste incorreto, tende a possuir medidas médias maiores que o método de Everitt/Paula. Este, por sua vez, sob o ajuste correto, demonstra ter medidas médias menores que o procedimento de Atkinson. Assim, de maneira geral, não há uma forma de obtenção do envelope que se destaque em relação a outra, considerando as duas avaliadas nesta análise.

Através da medida que avalia a proporção de vezes que pelo menos um ponto está fora do envelope, é possível verificar tanto por Atkinson quanto por Everitt/Paula que utilizar tal critério como forma de rejeição é ineficaz. Isso se deve ao fato da taxa de rejeição nula ser extremamente alta e crescente com o aumento do tamanho amostral. Para a taxa de rejeição não nula, notamos que em ambas as formas de obtenção do envelope a mesma tende a aumentar com o número de observações.

Por fim, através das taxas de rejeição do procedimento proposto, notamos que a taxa de rejeição nula parece flutuar de 0,04 a 0,06 para o gráfico normal e de 0,06 a 0,08 para o gráfico meio normal, para todos os cenários e tamanhos de amostra averiguados. Já a taxa de rejeição não nula parece ser próxima entre os gráficos de probabilidade normal e meio normal se a distribuição é a gama e consideravelmente maior no gráfico normal se a distribuição é a normal inversa. Podemos concluir que, na classe dos MLG com variável resposta contínua, tais apontamentos sugerem que é melhor usar nosso procedimento com o gráfico normal, pois o mesmo apresenta resultados semelhantes para uma distribuição e substancialmente melhores para a outra.

Tabela 5.13: Resultados de simulação para o cenário I-a - Estatísticas descritivas para algumas medidas - Variável resposta com distribuição normal inversa.

Método	Medida	Gráfico	Normal						Meio normal					
		n	50		100		200		50		100		200	
		Distribuição ajustada	NI	GA	NI	GA	NI	GA	NI	GA	NI	GA	NI	GA
		Taxa rejeição proc. prop.	**0,051**	**0,430**	**0,046**	**0,718**	**0,050**	**0,952**	0,068	0,119	0,069	0,209	0,076	0,408
Atkinson	Prop. de vezes de pelo menos um ponto fora		0,854	0,976	0,941	0,999	0,969	1,000	0,832	0,876	0,906	0,945	0,953	0,975
	Prop. pontos fora	Média	**0,089**	**0,232**	**0,091**	**0,320**	**0,094**	**0,475**	0,096	0,120	0,099	0,120	0,102	0,136
		D.P.	0,086	0,150	0,082	0,166	0,084	0,157	0,104	0,118	0,101	0,114	0,104	0,123
	Maior dist. ponto envel.	Média	0,095	0,309	0,101	0,501	0,101	0,760	0,068	0,090	0,066	0,161	0,068	0,302
		D.P.	0,106	0,290	0,105	0,425	0,119	0,557	0,094	0,126	0,085	0,265	0,100	0,458
	Prop. pontos consec. fora	Média	**0,051**	**0,123**	**0,048**	**0,162**	**0,044**	**0,247**	0,059	0,076	0,056	0,065	0,052	0,068
		D.P.	0,048	0,095	0,046	0,112	0,042	0,128	0,066	0,080	0,063	0,068	0,060	0,074
Everitt/ Paula	Prop. de vezes de pelo menos um ponto fora		0,615	0,903	0,716	0,985	0,783	1,000	0,582	0,630	0,662	0,737	0,742	0,875
	Prop. pontos fora	Média	**0,043**	**0,156**	**0,043**	**0,244**	**0,042**	**0,420**	0,050	0,062	0,046	0,063	0,047	0,081
		D.P.	0,061	0,133	0,059	0,156	0,056	0,158	0,078	0,092	0,072	0,085	0,072	0,099
	Maior dist. ponto envel.	Média	0,049	0,218	0,051	0,392	0,052	0,684	0,035	0,044	0,032	0,093	0,032	0,243
		D.P.	0,076	0,249	0,079	0,368	0,078	0,508	0,065	0,088	0,061	0,199	0,060	0,402
	Prop. pontos consec. fora	Média	**0,030**	**0,088**	**0,027**	**0,129**	**0,024**	**0,229**	0,035	0,044	0,031	0,040	0,028	0,045
		D.P.	0,040	0,082	0,037	0,102	0,032	0,125	0,053	0,064	0,049	0,057	0,044	0,061

Tabela 5.14: Resultados de simulação para o cenário I-b - Estatísticas descritivas para algumas medidas - Variável resposta com distribuição gama.

Método	Medida	Gráfico	Normal						Meio normal					
		n	50		100		200		50		100		200	
		Distribuição ajustada	GA	NI	GA	NI	GA	NI	GA	NI	GA	NI	GA	NI
		Taxa rejeição proc. prop.	0,036	0,459	0,037	0,776	**0,051**	**0,964**	0,075	0,512	0,073	0,778	0,074	0,961
Atkinson	Prop. de vezes de pelo menos um ponto fora		0,885	0,966	0,948	0,999	0,980	1,000	0,848	0,940	0,890	0,992	0,949	1,000
	Prop. pontos fora	Média	0,099	0,223	0,097	0,376	0,098	0,542	0,100	0,262	0,097	0,439	0,101	0,635
		D.P.	0,090	0,169	0,082	0,180	0,082	0,146	0,107	0,229	0,103	0,247	0,105	0,209
	Maior dist. ponto envel.	Média	0,114	0,464	0,114	0,715	0,108	1,102	0,070	0,433	0,067	0,681	0,068	1,070
		D.P.	0,124	0,572	0,122	0,680	0,125	0,947	0,092	0,581	0,091	0,686	0,099	0,954
	Prop. pontos consec. fora	Média	0,055	0,133	0,047	0,236	0,043	0,376	0,063	0,174	0,054	0,301	0,052	0,486
		D.P.	0,050	0,123	0,041	0,151	0,038	0,144	0,074	0,185	0,064	0,228	0,061	0,238
Everitt/ Paula	Prop. de vezes de pelo menos um ponto fora		0,653	0,858	0,764	0,985	0,828	1,000	0,560	0,825	0,6925	0,964	0,742	0,998
	Prop. pontos fora	Média	0,046	0,155	0,048	0,309	0,047	0,495	0,046	0,192	0,049	0,371	0,049	0,590
		D.P.	0,060	0,154	0,062	0,186	0,059	0,149	0,074	0,211	0,072	0,261	0,077	0,216
	Maior dist. ponto envel.	Média	0,057	0,338	0,062	0,627	0,057	0,974	0,033	0,317	0,035	0,603	0,034	0,950
		D.P.	0,096	0,494	0,090	0,696	0,095	0,824	0,068	0,498	0,064	0,702	0,074	0,829
	Prop. pontos consec. fora	Média	0,030	0,100	0,028	0,209	0,025	0,364	0,032	0,134	0,032	0,268	0,029	0,466
		D.P.	0,037	0,112	0,036	0,154	0,032	0,143	0,049	0,170	0,048	0,234	0,049	0,232

Os resultados referentes ao cenário I para as distribuições poisson e binomial negativa estão presentes na Tabela 5.15 e nas Tabelas A.11 a A.12 para os demais cenários (apresentadas no Apêndice A). Observa-se que em geral, sob o modelo correto ($PO \rightarrow PO$), a média das três medidas avaliadas flutuam em torno dos mesmos patamares que os MLG para distribuições contínuas, em ambas as formas de construção do envelope. Para o ajuste incorreto ($BN \rightarrow PO$), verificamos que as médias das quantidades computadas também tendem a ser maiores conforme aumenta-se o tamanho da amostra.

Comparando os gráficos de probabilidade normal e meio normal, notamos que os resultados são análogos a simulação em que geramos a variável resposta da distribuição gama, ou seja, não há um gráfico de probabilidade que se destaque em relação ao outro. No que se refere às formas de construção do envelope segundo Atkinson e Everitt/Paula, os aspectos constatados também são similares aos MLG para variáveis respostas contínuas.

Com base na proporção de vezes que pelo menos um ponto está fora do envelope, segundo as formas de Atkinson e Everitt/Paula para as distribuições poisson e binomial negativa, observa-se que os resultados fornecem interpretações semelhantes as realizadas aos MLG para distribuições contínuas. Por fim, avaliando as taxas de rejeição do método proposto, verificamos que nos três cenários a taxa de rejeição nula parece oscilar de 0,06 a 0,07 para o gráfico normal e de 0,09 a 0,12 para o gráfico meio normal. No que concerne a taxa de rejeição não nula, a mesma demonstra ser um pouco superior para o gráfico meio normal e tende a crescer com o aumento do tamanho amostral em ambos os gráficos. Podemos concluir que, assim como verificado para as distribuições normal inversa e gama, também é preferível utilizar o método proposto com o gráfico de probabilidade normal.

Tabela 5.15: Resultados de simulação para o cenário I - Estatísticas descritivas para algumas medidas - Variável resposta com distribuição poisson e binomial negativa.

		Gráfico	Normal						Meio normal					
		n	50		100		200		50		100		200	
Método	Medida	Distribuição gerada	PO	BN	PO	BN	PO	BN	PO	BN	PO	BN	PO	BN
		Taxa rejeição proc. prop.	0,069	0,477	0,071	0,716	0,058	0,938	0,104	0,548	0,105	0,780	0,100	0,962
Atkinson	Prop. de vezes de pelo menos um ponto fora		0,846	0,965	0,907	0,996	0,945	1,000	0,734	0,934	0,826	0,987	0,871	1,000
	Prop. pontos fora	Média	0,101	0,265	0,101	0,368	0,098	0,536	0,101	0,309	**0,099**	**0,445**	0,099	0,633
		D.P.	0,105	0,189	0,096	0,204	0,093	0,188	0,136	0,247	0,125	0,257	0,125	0,223
	Maior dist. ponto envel.	Média	0,124	0,386	0,121	0,449	0,117	0,524	0,074	0,297	0,072	0,348	0,072	0,423
		D.P.	0,143	0,323	0,131	0,315	0,140	0,318	0,113	0,280	0,100	0,267	0,111	0,266
	Prop. pontos consec. fora	Média	0,058	0,130	0,053	0,171	0,046	0,256	0,068	0,216	0,060	0,314	0,055	0,491
		D.P.	0,057	0,100	0,050	0,117	0,045	0,128	0,097	0,206	0,082	0,239	0,078	0,244
Everitt/ Paula	Prop. de vezes de pelo menos um ponto fora		0,590	0,887	0,671	0,978	0,754	0,999	0,462	0,837	0,553	0,956	0,619	0,998
	Prop. pontos fora	Média	0,050	0,193	0,047	0,299	0,046	0,470	0,050	0,235	**0,046**	**0,378**	0,047	0,574
		D.P.	0,076	0,177	0,069	0,207	0,068	0,202	0,100	0,233	0,087	0,267	0,090	0,236
	Maior dist. ponto envel.	Média	0,062	0,277	0,063	0,356	0,056	0,439	0,035	0,213	0,036	0,281	0,031	0,357
		D.P.	0,107	0,283	0,111	0,272	0,091	0,287	0,082	0,233	0,084	0,230	0,065	0,240
	Prop. pontos consec. fora	Média	0,033	0,101	0,028	0,146	0,026	0,231	0,037	0,170	0,030	0,278	0,029	0,458
		D.P.	0,046	0,093	0,039	0,113	0,037	0,129	0,077	0,190	0,057	0,241	0,058	0,251

5.2 Avaliação numérica: Modelos de regressão para taxas e proporções

Para os modelos de regressão para taxas e proporções, utilizamos para o estudo as distribuições beta e beta retangular, em que também definimos seis cenários. Nos cinco primeiros cenários atribuímos ao preditor a função de ligação logito e para o último, consideramos a função de ligação complemento log-log. Assim como nos MLG, trabalhamos com duas covariáveis. Para os quatro primeiros cenários e o último, geramos as variáveis preditoras de duas uniformes padrão independentes. Já para o cenário V, utilizamos para x_{i1} uma distribuição normal ($x_{i1} \sim N(0{,}5; 0{,}25^2)$) e para x_{i2} uma distribuição exponencial ($x_{i2} \sim Exp(0{,}3)$). É importante ressaltar, novamente, que as covariáveis foram geradas e depois mantidas fixas em todas as réplicas de Monte Carlo realizadas. Além disso, empregamos o parâmetro de mistura $\lambda = 0{,}7$ para a geração da variável resposta com distribuição beta retangular. Os cenários III e IV são aqueles presentes em Pereira (2017) e originalmente utilizados por Anholeto et al. (2014).

Para auxiliar a compreensão dos cenários, apresentamos na Tabela 5.16 a descrição dos mesmos. Para o cenário I, atribuímos ao vetor de parâmetros os valores $\boldsymbol{\beta} = (1; -0{,}5; 1{,}5)$ e desta maneira, $\mu \in (0{,}622; 0{,}924)$. No cenário II consideramos o parâmetro $\sigma = 0{,}3$ e nos demais $\sigma = 0{,}1$ a fim de avaliar os efeitos dessa modificação na performance dos métodos.

Tabela 5.16: Descrição dos cenários para o modelo de regressão beta e beta retangular.

Cenário	Função de ligação	Coeficientes	Covariáveis	μ	σ
I	$logito(\mu_i)$	$\beta_0 = 1, \beta_1 = -0,5, \beta_2 = 1,5$	$x_{i1} \sim U(0,1)$ e $x_{i2} \sim U(0,1)$	(0,622; 0,924)	0,1
II	$logito(\mu_i)$	$\beta_0 = 1, \beta_1 = -0,5, \beta_2 = 1,5$	$x_{i1} \sim U(0,1)$ e $x_{i2} \sim U(0,1)$	(0,622; 0,924)	0,3
III	$logito(\mu_i)$	$\beta_0 = -2,3, \beta_1 = -1,1, \beta_2 = -0,7$	$x_{i1} \sim U(0,1)$ e $x_{i2} \sim U(0,1)$	(0,016; 0,091)	0,1
IV	$logito(\mu_i)$	$\beta_0 = -0,3, \beta_1 = 0,7, \beta_2 = 0,3$	$x_{i1} \sim U(0,1)$ e $x_{i2} \sim U(0,1)$	(0,426; 0,668)	0,1
V	$logito(\mu_i)$	$\beta_0 = 1, \beta_1 = -0,5, \beta_2 = 1,5$	$x_{i1} \sim N(0,5; 0,25^2)$ e $x_{i2} \sim Exp(0,3)$	(0,609; 0,995)	0,1
VI	$cloglog(\mu_i)$	$\beta_0 = 0,4, \beta_1 = -0,5, \beta_2 = 0,7$	$x_{i1} \sim U(0,1)$ e $x_{i2} \sim U(0,1)$	(0,595; 0,950)	0,1

5.2.1 Comparação entre os procedimentos de rejeição do modelo

Similarmente ao que foi feito aos MLG, comparamos, para os modelos de regressão para taxas e proporções, o procedimento proposto com os métodos de rejeição baseados na construção do envelope segundo Flack e Flores e Lee e Rhee. Entretanto, avaliamos neste estudo duas situações:

- $BE \rightarrow BE$ representa a ocasião em que geramos a amostra de uma distribuição beta e construímos o envelope considerando uma distribuição beta;
- $BR \rightarrow BE$ representa a ocasião em que geramos a amostra de uma distribuição beta retangular e construímos o envelope considerando uma distribuição beta.

As especificações feitas aos modelos de regressão para taxas e proporções quanto a quantidade de réplicas de cada procedimento de rejeição e o valor de k utilizado nos métodos de Flack e Flores e Lee e Rhee foram as mesmas empregadas aos MLG. A única

modificação foi referente ao resíduo, em que para este caso utilizamos o resíduo quantílico para a obtenção das taxas de rejeição nula e não nula dos três procedimentos.

As Tabelas 5.17 a 5.22 apresentam os resultados contendo as taxas de rejeição do modelo referentes aos cenários de I a VI para os três métodos. Os aspectos gerais observados, quanto à taxa de rejeição nula ($BE \rightarrow BE$), a taxa de rejeição não nula ($BR \rightarrow BE$) e relacionados a características particulares dos três procedimentos, são similares aos vistos nos modelos lineares generalizados para as distribuições contínuas. Em particular, também observou-se que o procedimento proposto apresenta taxa de rejeição nula flutuando de 0,04 a 0,06, enquanto que, nos outros métodos, essa taxa cresce com o tamanho amostral. É interessante ressaltar que a performance relativa do nosso método em relação aos outros parece ser ainda melhor do que no caso do MLG com variável resposta contínua. Temos que para n = 50, nosso método apresenta performance superior ao de Flack e Flores e não inferior ao de Lee e Rhee. Já para n = 200, com exceção do cenário III, esse comportamento se inverte, pois notamos que procedimento proposto apresenta performance superior a de Lee e Rhee e não inferior a de Flack e Flores. Além disso, as modificações feitas quanto ao parâmetro σ (cenário II), ao suporte do preditor (cenários III e IV), a distribuição das variáveis preditoras (cenário V) e a função de ligação (cenário VI) demonstram influenciar apenas a taxa de rejeição não nula dos métodos.

Tabela 5.17: Resultados de simulação para o cenário I - Taxas de rejeição nula e rejeição não nula.

n	Método	$BE \rightarrow BE$	$BR \rightarrow BE$
50	Flack e Flores	0,048	0,362
	Lee e Rhee	0,034	0,297
	Proposto	**0,041**	**0,522**
100	Flack e Flores	0,064	0,854
	Lee e Rhee	0,052	0,762
	Proposto	**0,039**	**0,885**
200	Flack e Flores	0,084	0,994
	Lee e Rhee	0,064	0,983
	Proposto	**0,053**	**0,994**

Tabela 5.18: Resultados de simulação para o cenário II - Taxas de rejeição nula e rejeição não nula.

n	Método	$BE \rightarrow BE$	$BR \rightarrow BE$
50	Flack e Flores	0,054	0,132
	Lee e Rhee	0,030	0,085
	Proposto	**0,043**	**0,176**
100	Flack e Flores	0,068	0,304
	Lee e Rhee	0,054	0,198
	Proposto	**0,041**	**0,361**
200	Flack e Flores	0,092	0,652
	Lee e Rhee	0,066	0,515
	Proposto	**0,054**	**0,665**

Tabela 5.19: Resultados de simulação para o cenário III - Taxas de rejeição nula e rejeição não nula.

n	Método	$BE \rightarrow BE$	$BR \rightarrow BE$
50	Flack e Flores	0,050	0,314
	Lee e Rhee	0,035	0,242
	Proposto	**0,037**	**0,437**
100	Flack e Flores	0,070	0,850
	Lee e Rhee	0,052	0,829
	Proposto	0,037	0,814
200	Flack e Flores	0,094	1,000
	Lee e Rhee	0,065	0,999
	Proposto	0,054	0,986

Tabela 5.20: Resultados de simulação para o cenário IV - Taxas de rejeição nula e rejeição não nula.

n	Método	$BE \rightarrow BE$	$BR \rightarrow BE$
50	Flack e Flores	0,056	0,406
	Lee e Rhee	0,031	0,321
	Proposto	**0,040**	**0,408**
100	Flack e Flores	0,066	0,780
	Lee e Rhee	0,054	0,711
	Proposto	**0,039**	**0,760**
200	Flack e Flores	0,092	0,982
	Lee e Rhee	0,063	0,965
	Proposto	**0,052**	**0,983**

Tabela 5.21: Resultados de simulação para o cenário V - Taxas de rejeição nula e rejeição não nula.

n	Método	$BE \rightarrow BE$	$BR \rightarrow BE$
50	Flack e Flores	0,060	0,434
	Lee e Rhee	0,027	0,361
	Proposto	**0,051**	**0,556**
100	Flack e Flores	0,070	0,812
	Lee e Rhee	0,052	0,734
	Proposto	**0,056**	**0,821**
200	Flack e Flores	0,102	1,000
	Lee e Rhee	0,068	0,988
	Proposto	**0,046**	**0,992**

Tabela 5.22: Resultados de simulação para o cenário VI - Taxas de rejeição nula e rejeição não nula.

n	Método	$BE \rightarrow BE$	$BR \rightarrow BE$
50	Flack e Flores	0,054	0,354
	Lee e Rhee	0,032	0,276
	Proposto	**0,037**	**0,498**
100	Flack e Flores	0,062	0,816
	Lee e Rhee	0,052	0,724
	Proposto	**0,041**	**0,853**
200	Flack e Flores	0,092	0,988
	Lee e Rhee	0,061	0,979
	Proposto	**0,051**	**0,991**

5.2.2 Comparação entre o gráfico normal e meio normal probabilístico com envelope simulado

Para a comparação dos gráficos de probabilidade nos modelos de regressão para taxas de proporções, também realizamos 2000 réplicas via Monte Carlo do processo de geração do envelope simulado para os métodos de construção segundo Atkinson e Everitt/Paula. As estatísticas obtidas a partir do envelope foram as mesmas que nos MLG. Entretanto, utilizamos nesta classe de modelos o resíduo quantílico para obtê-las.

Os resultados referentes ao cenário I estão presentes na Tabela 5.23 e nas Tabelas de B.1 a B.5 para os demais cenários (apresentadas no Apêndice B). De maneira geral, vemos que sob o modelo correto ($BE \rightarrow BE$), as médias da proporção de pontos fora e da proporção de pontos consecutivos fora do envelope oscilam em torno dos mesmos patamares que nos MLG, tanto para o procedimento de construção segundo Atkinson quanto conforme Everitt/Paula. Já sob o ajuste incorreto, $BR \rightarrow BE$, vemos que as médias das medidas tendem a ser maiores conforme aumenta-se a quantidade de observações. No que concerne as formas de construção do envelope (Atkinson e Everitt/Paula), obtemos para os modelos de regressão para taxas e proporções as mesmas conclusões apresentadas aos modelos lineares generalizados.

Comparando os gráficos de probabilidade normal e meio normal, verificamos que nos cenários I, II, V e VI, o gráfico normal apresenta melhor performance que o gráfico de probabilidade meio normal em ambas as formas de construção do envelope, com base na média da proporção de pontos consecutivos fora do envelope. É interessante notar que sob o ajuste incorreto, o gráfico normal possui um desempenho relativamente melhor que o gráfico meio normal, considerando a média das três medidas avaliadas.

Para os cenários III e IV, observamos resultados parecidos entre si e diferentes dos demais cenários. Em ambos não houve um gráfico de probabilidade que se destacasse em pelo menos uma das medidas avaliadas nas formas de construção do envelope segundo Atkinson e Everitt/Paula. Entretanto, de maneira geral, o gráfico de probabilidade meio normal com envelope apresentou um desempenho melhor na indicação do ajuste incorreto, conforme apontado pelas médias da proporção de pontos fora e a proporção de pontos consecutivos fora do envelope. Vale notar que nesses cenários, a modificação foi realizada no intervalo de variação da média, em que o mesmo está próximo de zero no cenário III e em torno de 0,5 no cenário IV.

Assim como verificado na classe dos MLG, a métrica que avalia a proporção de vezes que pelo menos um ponto está fora do envelope tende a crescer com o tamanho da amostra e apresenta valores elevados da proporção tanto sob o modelo correto quanto considerando o ajuste incorreto. Novamente, esse aspecto reforça a ineficácia de tal critério como condição para rejeição de uma dada distribuição assumida para a variável resposta.

Com base nas taxas de rejeição do método proposto, verificamos que em todos os cenários a taxa de rejeição nula parece oscilar de 0,04 a 0,05 para o gráfico normal e de 0,07

a 0,08 para o gráfico meio normal, independentemente do tamanho amostral. Já para a taxa de rejeição não nula, a mesma parece ser próxima entre os gráficos de probabilidade normal e meio normal nos cenários III e IV e substancialmente maior com o gráfico normal nos cenários I, II, V e VI. Por meio de tais apontamentos, podemos concluir que assim como nos MLG, é preferível utilizar o método proposto com o gráfico de probabilidade normal nesta classe de modelos.

Tabela 5.23: Resultados de simulação para o cenário I - Estatísticas descritivas para algumas medidas - Variável resposta com distribuição beta e beta retangular.

		Gráfico	Normal						Meio normal					
		n	50		100		200		50		100		200	
Método	Medida	Distribuição gerada	BE	BR	BE	BR	BE	BR	BE	BR	BE	BR	BE	BR
		Taxa rejeição proc. prop.	**0,041**	**0,522**	**0,039**	**0,885**	**0,053**	**0,994**	0,070	0,204	0,070	0,421	0,074	0,723
Atkinson	Prop. de vezes de pelo menos um ponto fora		0,892	0,982	0,953	0,999	0,986	1,000	0,835	0,900	0,901	0,976	0,949	0,998
	Prop. pontos fora	Média	0,102	0,288	0,098	0,451	0,101	0,591	0,101	0,154	0,097	0,242	0,098	0,374
		D.P.	0,090	0,165	0,085	0,155	0,083	0,114	0,107	0,140	0,105	0,166	0,100	0,174
	Maior dist. ponto envel.	Média	0,111	0,234	0,112	0,289	0,108	0,328	0,065	0,101	0,065	0,120	0,067	0,134
		D.P.	0,119	0,172	0,122	0,159	0,123	0,158	0,088	0,132	0,086	0,137	0,098	0,146
	Prop. pontos consec. fora	Média	**0,055**	**0,161**	**0,047**	**0,260**	**0,043**	**0,356**	0,062	0,098	0,054	0,147	0,050	0,223
		D.P.	0,049	0,109	0,042	0,114	0,038	0,083	0,070	0,100	0,063	0,120	0,059	0,134
Everitt/ Paula	Prop. de vezes de pelo menos um ponto fora		0,662	0,936	0,774	0,997	0,829	1,000	0,549	0,719	0,691	0,899	0,750	0,980
	Prop. pontos fora	Média	0,047	0,225	**0,050**	**0,392**	**0,048**	**0,547**	0,047	0,091	0,052	0,170	0,049	0,305
		D.P.	0,061	0,161	0,063	0,152	0,060	0,113	0,076	0,110	0,079	0,150	0,074	0,171
	Maior dist. ponto envel.	Média	0,056	0,172	0,063	0,236	0,055	0,279	0,032	0,061	0,035	0,082	0,032	0,096
		D.P.	0,092	0,141	0,093	0,138	0,096	0,119	0,067	0,099	0,064	0,110	0,074	0,101
	Prop. pontos consec. fora	Média	**0,031**	**0,136**	**0,029**	**0,243**	**0,026**	**0,345**	0,033	0,064	0,033	0,111	0,029	0,196
		D.P.	0,036	0,106	0,036	0,109	0,032	0,078	0,054	0,081	0,053	0,111	0,046	0,131

Capítulo 6

Aplicação

Neste capítulo é apresentada a aplicação dos procedimentos de rejeição baseados em envelopes simulados a dados referentes a Pesquisa Nacional de Saúde (PNS) realizada em 2013. Na Seção 6.1 é feita uma breve introdução e descrição do conjunto de dados utilizado, além da análise descritiva do mesmo. O ajuste do modelo e a interpretação dos resultados estão presentes na Seção 6.2, enquanto a análise de diagnóstico e discussão dos métodos de rejeição são apresentadas na Seção 6.3.

6.1 Descrição dos dados

O conjunto de dados é proveniente da primeira edição da Pesquisa Nacional da Saúde (PNS) realizada em 2013 pelo Instituto Brasileiro de Geografia e Estatística (IBGE) em convênio com o Ministério da Saúde. O intuito da pesquisa está em avaliar em âmbito nacional aspectos relacionados à saúde e estilo de vida da população brasileira. Para isso, foram coletados dados de 205.546 indivíduos divididos em 26 estados do país, além do Distrito Federal, os quais estão disponíveis em IBGE (2018).

O objetivo desta aplicação é identificar quais hábitos alimentares, além de fatores inerentes (idade e gênero), influenciam na condição de obesidade abdominal e consequentemente, no risco do desenvolvimento de doenças cardiovasculares, diabetes, hipertensão e outras em indivíduos do Rio Grande do Sul (RS).

A obesidade é apontada pela Organização Mundial de Saúde (OMS) como um dos maiores problemas de saúde pública no mundo (ABESO, 2018). Segundo pesquisa realizada pelo Ministério da Saúde, aproximadamente metade da população brasileira está acima do peso. Além disso, Porto Alegre (RS) é a capital do país que possui a maior quantidade de pessoas com excesso de peso, com cerca de 55,4% (SBEM, 2018).

No que diz respeito a obesidade abdominal, ela se caracteriza pelo acúmulo de gordura na região do abdômen, a qual leva a um aumento do risco de certas doenças crônicas, tais como doenças cardiovasculares, diabetes, hipertensão e colesterol alto. Estes riscos existem pois há uma grande quantidade de gordura presente na região visceral, ou seja,

próxima ao coração (MS, 2018). Para averiguar a condição de obesidade abdominal, uma medida utilizada é a circunferência da cintura, visto que esta é mais fortemente correlacionada com o tecido adiposo visceral do que a circunferência medida no nível do umbigo (Willis et al., 2007).

Trabalharemos no conjunto de dados com os entrevistados do Rio Grande do Sul (RS) maiores de 18 anos que realizaram exames laboratoriais, resultando em um total de 2874 indivíduos. Além disso, consideraremos para análise 13 variáveis, sendo duas a respeito de fatores inerentes ao indivíduo, tais como idade e gênero, além de onze variáveis alimentares, as quais são referentes ao consumo semanal de feijão, legume cru, legume cozido, carne vermelha, frango, peixe, suco natural, frutas, refrigerante/suco artificial, leite e bebida alcoólica.

Inicialmente, realizamos a análise descritiva avaliando o comportamento da circunferência da cintura dos indivíduos, conforme Figura 6.1 e Tabela 6.1. É possível observar uma grande concentração de medidas da circunferência da cintura em torno de 90 cm. Além disso, a distribuição de probabilidade da circunferência da cintura parece ser ligeiramente assimétrica à direita.

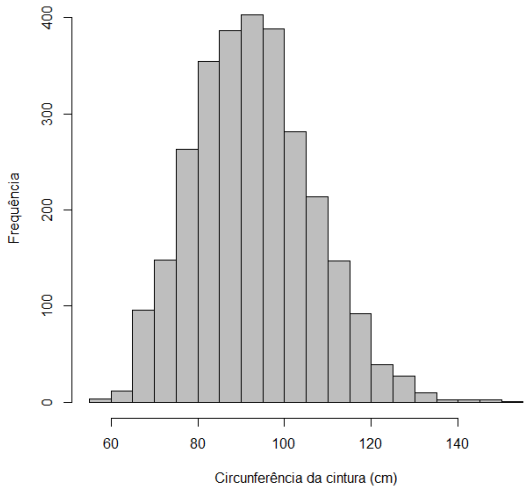

Figura 6.1: Histograma da circunferência da cintura dos indivíduos do RS.

Tabela 6.1: Medidas descritivas da variável circunferência da cintura.

Variável	N	Média	D.P.	Mín	Q1	Mediana	Q3	Máx
Circunf. da cintura	2874	93,02	13,79	57,80	83,22	92,30	101,70	152,00

Prosseguindo a análise descritiva, avaliaremos agora o comportamento das covariáveis consideradas no estudo. Na Figura 6.2, apresentamos a relação entre a circunferência da cintura do indivíduo e sua respectiva idade. Observe que devido a alta variabilidade presente nos dados, na Figura 6.2 (esquerda) não é possível perceber nenhuma relação entre ambas as variáveis. Desta forma, optou-se por obter para cada idade a respectiva média da circunferência da cintura, cujo comportamento pode ser visto na Figura 6.2 (direita).

Por meio da Figura 6.2 (direita), verifica-se aparentemente um comportamento quadrático da circunferência da cintura média conforme aumenta-se a idade dos indivíduos. Além disso, observa-se uma alta variabilidade para idades mais elevadas, em virtude de um menor número de observações de pessoas idosas.

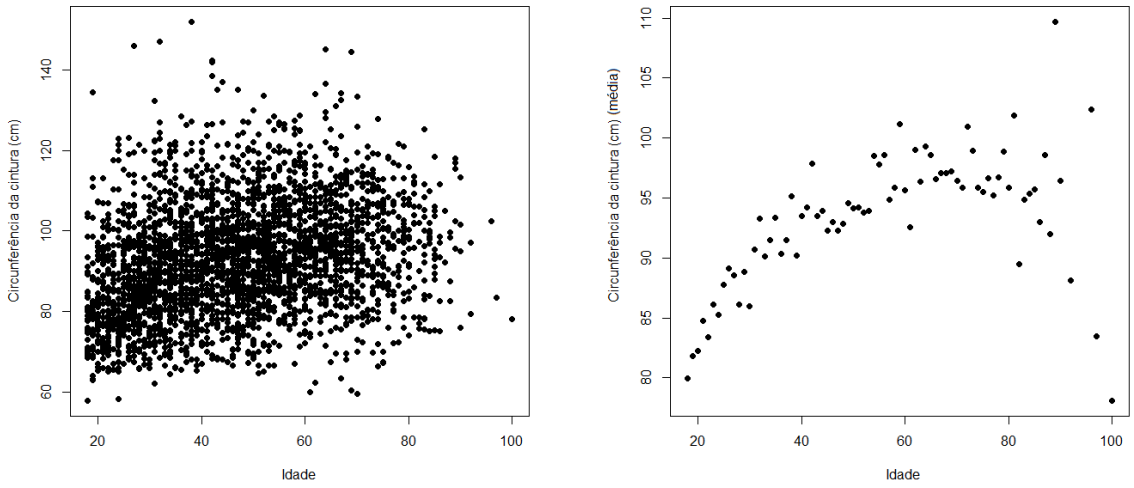

Figura 6.2: Gráfico de dispersão da circunferência da cintura pela idade dos indivíduos (esquerda) e gráfico de dispersão da média da circunferência da cintura para cada idade dos indivíduos (direita).

Outra variável considerada na análise é o gênero, cujas medidas descritivas da circunferência da cintura segmentada pelas categorias feminino e masculino estão presentes na Tabela 6.2. Observa-se que em média, a circunferência da cintura dos homens é aproximadamente 5,5 cm maior do que das mulheres entre os indivíduos, considerando os entrevistados do Rio Grande do Sul.

Tabela 6.2: Medidas descritivas da circunferência da cintura segmentada pelas categorias de gênero.

	Categoria	%	Média	D.P.	Mín	Q1	Mediana	Q3	Máx
Gênero	Masculino	44,54	96,1	13,1	57,8	87,2	95,0	105,0	152,0
	Feminino	55,46	90,5	13,8	58,3	80,5	89,5	99,4	147,0

As demais covariáveis estudadas foram obtidas no questionário da PNS de forma categórica e, desta maneira, definimos a categorização final por meio de análises descritivas prévias. Na Tabela 6.3 apresentamos as medidas descritivas da circunferência da cintura segmentada pelas variáveis alimentares.

Com base nos valores médios das categorias, observa-se que aparentemente as variáveis associadas ao consumo de carne vermelha, frango, peixe, suco natural, refrigerante/-suco artificial, além de bebida alcoólica parecem apresentar médias da circunferência da cintura diferentes entre suas categorias. De forma mais específica, ao que parece, quanto maior é o consumo de carne vermelha, refrigerante/suco artificial e bebida alcoólica, maior

também é, em média, a circunferência abdominal do indivíduo. Em contraponto, temos que, aparentemente, quanto mais uma pessoa consome peixe e suco natural, menor é, em média, a circunferência de sua cintura. Com relação ao frango, ao que tudo indica, é preferível comê-lo tirando a pele do que deixar com a pele ou simplesmente não se alimentar desse tipo de proteína. Além disso, não parece haver diferença na média da circunferência da cintura entre as categorias referentes ao consumo de feijão, legume cru ou cozido, frutas e leite.

Tabela 6.3: Medidas descritivas da circunferência da cintura segmentada pelas categorias de variáveis referentes a alimentação.

	Categoria	%	Média	D.P.	Mín	Q1	Mediana	Q3	Máx
Feijão	Até 5 dias	60,3	92,8	14,2	60,0	82,5	92,0	101,7	147,0
	Mais de 5 dias	39,7	93,4	13,1	57,8	84,3	93,0	101,7	152,0
Legume cru	Até 5 dias	45,6	92,5	14,1	57,8	82,3	92,0	101,5	152,0
	Mais de 5 dias	54,4	93,4	13,5	60,0	84,0	92,8	102,0	147,0
Legume cozido	Até 5 dias	72,3	93,1	13,9	57,8	83,3	92,5	101,6	152,0
	Mais de 5 dias	27,7	92,7	13,6	63,0	83,0	91,8	101,8	147,0
Carne vermelha	Nunca	3,5	89,4	14,1	64,5	77,5	88,2	98,0	127,3
	Até 2 dias	17,6	92,0	14,2	57,8	81,5	91,8	101,3	133,5
	3 dias ou + s/gord	56,8	92,7	13,7	58,3	83,3	92,0	101,3	152,0
	3 dias ou + c/gord	22,1	95,2	13,4	64,6	85,1	94,5	104,6	144,5
Frango	Nunca	5,7	93,5	14,1	59,6	83,0	93,8	103,5	127,3
	Tirando a pele	67,2	92,5	13,7	58,3	82,8	92,0	101,0	147,0
	Com a pele	27,1	94,3	13,9	57,8	84,3	93,3	103,5	152,0
Peixe	Até 2 dias	95,0	93,1	13,8	57,8	83,3	92,5	101,8	152,0
	3 dias ou +	5,0	91,7	13,8	65,0	82,0	90,2	99,0	128,6
Suco natural	Nunca	38,6	93,8	14,1	60,3	83,5	93,3	102,9	152,0
	1 a 5 dias	45,2	92,7	13,7	57,8	83,0	91,8	101,1	147,0
	Mais de 5 dias	16,2	92,1	13,3	59,6	83,0	92,0	101,0	134,5
Frutas	Até 5 dias	51,8	92,7	14,1	57,8	82,8	92,0	101,5	152,0
	Mais de 5 dias	48,2	93,4	13,5	60,0	83,5	92,8	102,1	145,2
Refrigerante Suco artificial	Até 5 dias	80,2	92,6	13,5	57,8	83,0	92,0	101,3	146,0
	Mais de 5 dias	19,8	94,6	14,6	63,1	84,3	94,0	103,6	152,0
Leite	Até 5 dias	46,1	93,1	13,9	57,8	83,3	92,3	101,9	147,0
	Mais de 5 dias	53,9	92,9	13,7	59,6	83,1	92,3	101,5	152,0
Bebida alcoólica	Nunca	67,9	92,7	13,9	57,8	83,0	92,1	101,5	147,0
	Até 1 dia	16,1	93,1	13,7	60,0	83,0	93,0	101,5	142,3
	2 dias ou +	16,0	94,3	13,5	65,3	84,5	92,3	103,3	152,0

6.2 Ajuste do modelo

Para o ajuste do modelo, foram consideradas todas as variáveis presentes na análise descritiva, sendo mantidas aquelas cuja inclusão pelo teste da razão de verossimilhanças foram significativas ao nível de 5% tanto no teste contendo todas as variáveis, quanto no modelo com apenas a mesma no preditor.

Após diversos ajustes e testes, o modelo final encontrado contém 5 variáveis preditoras. Na Tabela 6.4 está disponível o AIC dos modelos considerando as distribuições normal, gama e normal inversa para a circunferência da cintura.

Tabela 6.4: AIC para as diferentes distribuições.

Distribuição	**AIC**
Normal	22785,51
Gama	22700,78
Normal Inversa	22696,32

Observe que utilizando o AIC como critério informal para avaliar qual distribuição melhor se adequa aos dados da circunferência da cintura, verifica-se que a distribuição normal inversa é a que apresenta menor valor segundo esse critério e, desta maneira, a assumiremos para o estudo em questão.

O ajuste utilizando a distribuição normal inversa é apresentado na Tabela 6.5. Vale ressaltar que a função de ligação adotada é a logarítmica. Além disso, a variável idade está presente no ajuste de forma linear e quadrática, conforme evidenciado na análise descritiva.

Tabela 6.5: Estimativa dos parâmetros do modelo de regressão normal inverso.

Variável	Categoria	Estimativa	E.P.	Valor t	p-valor	exp()
Intercepto	—	4,26795	0,02306	185,096	<0,0001	—
Idade	Linear	0,00872	0,00080	10,942	<0,0001	1,00875
	Quadrático	-0,00006	0,00001	-7,678	<0,0001	0,99994
Gênero	Feminino	-0,06067	0,00527	-11,512	<0,0001	0,94113
Carne vermelha	Até 2 dias	0,03034	0,01436	2,113	0,0347	1,03080
	3 dias ou + s/gord	0,04126	0,01354	3,047	0,0023	1,04212
	3 dias ou + c/gord	0,05351	0,01432	3,737	0,0002	1,05497
Refri./Suco art.	Mais de 5 dias	0,02443	0,00667	3,666	0,0003	1,02473
Suco natural	1 a 5 dias	-0,00727	0,00563	-1,292	0,1965	0,99276
	Mais de 5 dias	-0,01481	0,00765	-1,937	0,0529	0,98530

Por meio da Tabela 6.5, podemos discutir os sinais das estimativas dos parâmetros presentes no modelo. Nota-se que a idade possui uma relação quadrática com a circunferência da cintura, ou seja, a circunferência da cintura média do indivíduo tende a ser maior até um certo patamar de idade e, após isso, para pessoas mais idosas, a circunferência da cintura média tende a diminuir. Quanto ao gênero, verifica-se que a circunferência da cintura média das mulheres é inferior a dos homens. No que se refere ao consumo de carne vermelha, constata-se que indivíduos que a ingerem em até 2 dias na semana já apresentam circunferência da cintura média superior aos que não consomem. E o tamanho médio da cintura ainda aumenta quando consideramos os que não ingerem com pessoas que consomem carne vermelha em 3 ou mais dias na semana. Com relação ao refrigerante e suco artificial, verifica-se que indivíduos que os consomem em mais de 5 dias na semana apresentam circunferência da cintura média superior aos que não ingerem esse tipo de produto ou ingerem em uma quantidade menor de dias. Já para aqueles que consomem suco natural em mais de 5 dias na semana, nota-se que os mesmos apresentam circunferência da cintura média inferior aos que não ingerem ou ingerem em menos dias.

Considerando novamente a Tabela 6.5, podemos ainda interpretar as estimativas dos parâmetros a partir do exponencial dos coeficientes estimados. Para o gênero, por exemplo, quando passamos de uma pessoa do sexo masculino para o feminino, estima-se que a média da circunferência da cintura reduz 5,89% (é multiplicada por 0,941), mantidas as demais variáveis explicativas constantes. Com relação ao consumo de carne vermelha, temos que ao passar de um indivíduo que não ingere para um que consome em 3 dias ou mais por semana sem gordura, estima-se que a média da circunferência da cintura aumente em 4,21%, mantidas as demais variáveis preditores constantes. A interpretação das demais estimativas pode ser feita de maneira análoga.

Por fim, vale ressaltar que, embora o consumo de peixe, frango e bebida alcoólica apresentem aparentes diferenças entre suas categorias, constatou-se que as mesmas não demonstraram ser significativas no modelo. Além disso, podemos verificar que os sinais das estimativas dos parâmetros está coerente com a análise descritiva realizada.

6.3 Análise de diagnóstico

Para a análise de diagnóstico, abordaremos primeiramente aspectos relacionados a suposição da distribuição assumida para a circunferência da cintura através dos métodos de rejeição expostos no Capítulo 4. Posteriormente, apresentaremos alguns estudos para identificação de pontos de alavanca, discrepantes e influentes. Em todas as análises utilizou-se o resíduo componente do desvio.

As Figuras de 6.3 a 6.5 trazem a aplicação dos métodos de rejeição: proposto, de Lee e Rhee e Flack e Flores, respectivamente, ao modelo obtido na Tabela 6.5. As figuras compreendidas à esquerda correspondem ao ajuste discutido na Seção 6.2 (distribuição normal inversa para a circunferência da cintura dadas as variáveis preditoras). Já as figuras dispostas à direita representam a aplicação dos procedimentos de rejeição se houvéssemos assumido para a variável resposta uma distribuição gama dadas as variáveis preditoras contidas na Tabela 6.5.

É possível avaliar que, sob o ajuste da distribuição normal inversa, apenas o procedimento proposto indicou a não rejeição da mesma. Esse fato possivelmente se deve, conforme visto nas simulações do Capítulo 5, porque os métodos baseados nas formas de construção de Lee e Rhee e Flack e Flores demonstram possuir a taxa de rejeição nula crescente quando se aumenta o tamanho da amostra. Já o procedimento proposto, como verificamos, não apresenta tal comportamento e, desta forma, tendo ele como base, não há indícios de inadequação do modelo obtido.

Considerando o modelo assumindo uma distribuição gama para a variável resposta, notamos que todos os três procedimentos são concordantes e recomendam a rejeição da referida distribuição. É importante ressaltar que para o método proposto, a rejeição ocorre segundo os dois critérios presentes em sua construção, ou seja, pela proporção de pontos fora e pela distância entre a banda do envelope e o ponto mais afastado desse limite.

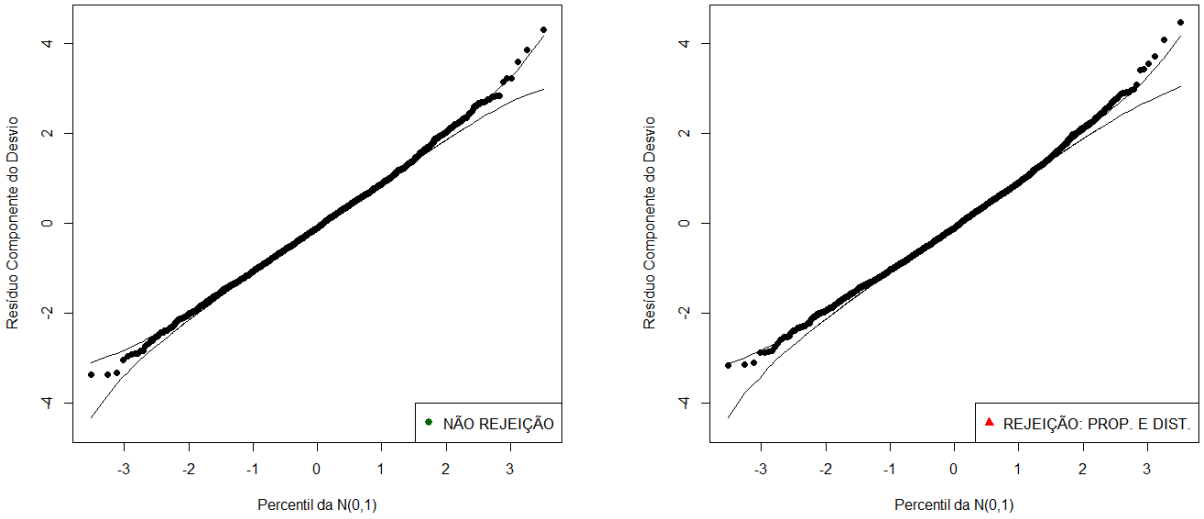

Figura 6.3: Método de rejeição proposto aplicado ao gráfico normal probabilístico utilizando a distribuição normal inversa (esquerda) e a distribuição gama (direita).

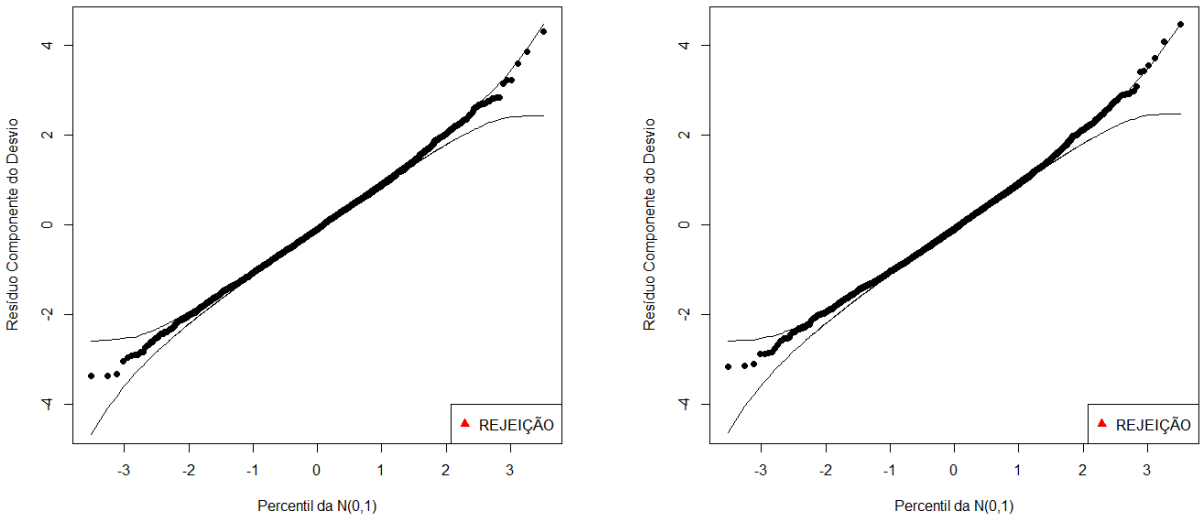

Figura 6.4: Método de rejeição do Lee e Rhee aplicado ao gráfico normal probabilístico utilizando a distribuição normal inversa (esquerda) e a distribuição gama (direita).

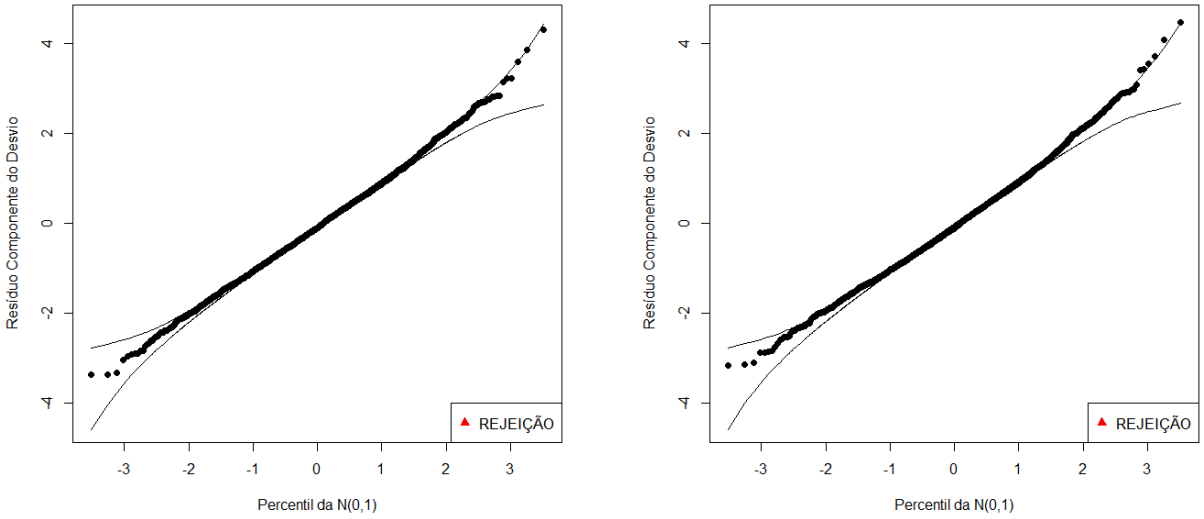

Figura 6.5: Método de rejeição do Flack e Flores aplicado ao gráfico normal probabilístico utilizando a distribuição normal inversa (esquerda) e a distribuição gama (direita).

Dando continuidade à análise de diagnóstico do modelo, o estudo e identificação de observações discrepantes, influentes e de alavanca também pode ser feito graficamente.

Primeiramente, abordaremos a identificação de observações discrepantes. Para a detecção de pontos aberrantes, usualmente utilizamos o gráfico contendo os valores ajustados pelo modelo no eixo horizontal e alguma forma de resíduo no eixo vertical. São consideradas discrepantes as observações que em módulo apresentam valores de resíduo muito superior aos demais. Através da Figura 6.6, é possível identificar que as observações 475, 1340 e 2532 são possíveis pontos aberrantes no ajuste. Apesar disso, os resíduos demonstram ter um comportamento aleatório ao redor do zero, com uma faixa de variação no intervalo $(-3{,}5; 3{,}5)$, e assim não há evidências de problemas relacionados com o ajuste do modelo.

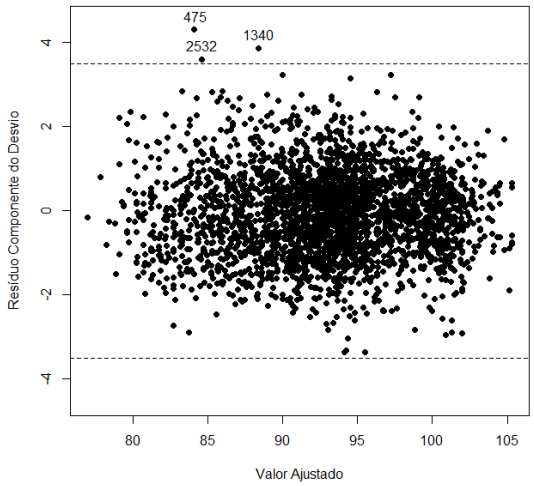

Figura 6.6: Gráfico dos valores ajustados pelo resíduo componente do desvio.

Com relação a observações influentes, verificamos tais ocorrências com o gráfico contendo a distância de Cook na ordenada e a ordem das observações na abcissa. Os pontos que se afastarem dos demais são aqueles que devem ser avaliados quanto a sua influência no ajuste do modelo. Pela Figura 6.7, constata-se que as observações 781, 2405 e 2532 aparecem como possíveis pontos influentes.

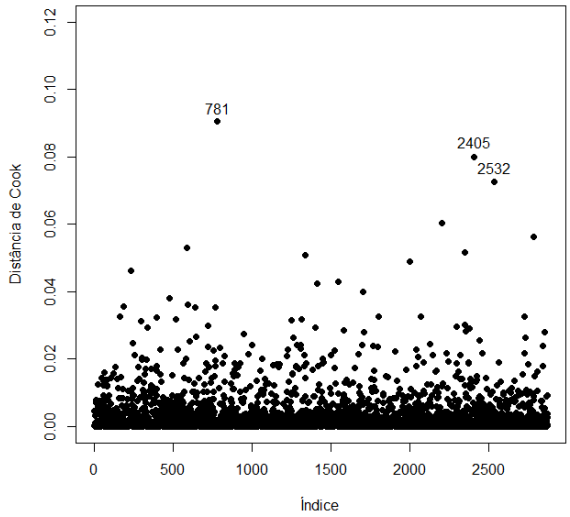

Figura 6.7: Gráfico da distância de Cook para as estimativas dos coeficientes.

Por fim, para identificar possíveis pontos de alavanca, construímos o gráfico contendo os valores da diagonal da matriz de projeção no eixo vertical e a ordem das observações no eixo horizontal, em que assim como no caso de observações influentes, os pontos considerados de alavanca serão aqueles que se afastarem relativamente dos demais. Com base na Figura 6.8, notamos que apenas a observação 280 aparece como possível ponto de alavanca, uma vez que apresenta medida h mais elevada.

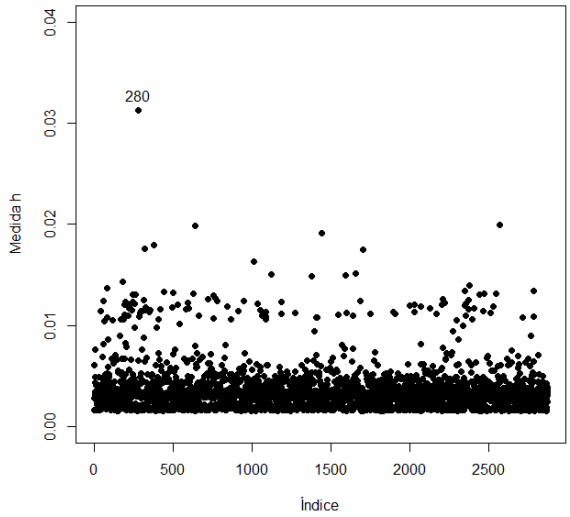

Figura 6.8: Gráfico da ordem das observações pelas medidas h.

Realizada a etapa de identificação de possíveis observações discrepantes, influentes e de alavanca, na Tabela 6.6 há o estudo da remoção individual e conjunta dos pontos, no intuito de avaliar o impacto dos mesmos no ajuste.

Observa-se que, de forma geral, tanto a retirada individual quanto conjunta das observações no modelo modificam pouco as estimativas dos coeficientes. Desta maneira, as conclusões discutidas até o momento não se alteram com a exclusão dos pontos identificados. Assim, temos mais um indício de que o modelo considerado é adequado para o ajuste da circunferência da cintura. Além disso, é interessante relembrar que o método proposto neste trabalho não rejeitou a adequabilidade deste modelo. Já os demais métodos, baseados nas propostas de Flack e Flores e Lee e Rhee, apresentaram decisão contrária, indicando a rejeição do ajuste.

Tabela 6.6: Variação percentual das estimativas dos parâmetros com a retirada das observações.

			Mudança % sem a observação						
Variável	Categoria	Estimativa	280	475	781	1340	2405	2532	Sem todas
Intercepto	—	4,26795	0,050	-0,035	-0,066	0,017	-0,096	-0,092	-0,227
Idade	Linear	0,00872	-0,784	0,779	-0,674	-0,022	-0,303	1,713	0,756
	Quadrático	-0,00006	-1,226	0,813	-1,058	-0,166	-0,584	2,116	-0,034
Gênero	Feminino	-0,06067	-0,023	0,859	0,357	0,879	0,308	-0,923	1,472
Carne vermelha	Até 2 dias	0,03034	-2,889	0,153	13,985	0,244	13,886	-0,399	25,540
	3 dias ou + s/gord	0,04126	-2,055	-1,019	10,238	-0,778	10,216	-1,040	15,961
	3 dias ou + c/gord	0,05351	-1,512	-0,143	7,836	0,217	7,865	0,015	14,583
Refri./Suco art.	Mais de 5 dias	0,02443	0,031	1,319	0,074	-5,860	1,115	1,212	-2,109
Suco natural	1 a 5 dias	-0,00727	-0,006	7,506	0,153	10,180	-6,026	-1,389	10,385
	Mais de 5 dias	-0,01481	-0,933	-0,306	6,006	2,297	-3,718	8,817	12,283

Capítulo 7

Conclusão

Neste trabalho, discorremos inicialmente a respeito de duas classes de modelos: os modelos lineares generalizados e os modelos de regressão para taxas e proporções. Posteriormente, discutimos acerca dos gráficos de probabilidade normal e meio normal, assim como apresentamos quatro formas de construção do envelope simulado, dentre elas segundo Atkinson, Everitt/Paula, Flack e Flores e Lee e Rhee. Após isso, propomos um método de rejeição do modelo a partir do envelope, o qual é baseado no método de construção conforme Atkinson. No intuito de fornecer uma base comparativa ao procedimento proposto, sugerimos um critério de rejeição do modelo para duas formas de obtenção do envelope (Flack e Flores e Lee e Rhee), em que a rejeição ocorre se pelo menos um ponto estiver fora do envelope.

Por meio de estudos de simulação via Monte Carlo, observamos que, em geral, o método proposto apresenta taxas estáveis de rejeição do modelo sob a distribuição correta, conforme aumenta-se o tamanho da amostra, principalmente sob as classes dos MLG para variáveis respostas contínuas e para os modelos de regressão para taxas e proporções. Já para as demais metodologias, além de possuírem um custo computacional maior, a taxa de rejeição do modelo correto cresce conforme aumenta-se o número de observações. Com relação a taxa de rejeição do ajuste incorreto, é possível notar que todos os procedimentos tendem a possuir a taxa de rejeição maior conforme aumenta-se a quantidade de observações. Isso significa que quanto maior o tamanho amostral, mais fácil torna-se para os métodos indicarem quando a distribuição incorreta é considerada.

Complementando os resultados, realizamos também a comparação do gráfico de probabilidade normal e meio normal com envelope simulado através de estudos de Monte Carlo. Para isso, consideramos os procedimentos de construção segundo Atkinson e Everitt/Paula, em que para cada uma das réplicas do processo de geração do envelope computamos estatísticas referentes a proporção de pontos fora do envelope, a distância entre a banda do envelope e o ponto mais afastado desse limite e a proporção máxima de pontos consecutivos além das bandas do envelope. Verificou-se que, de maneira geral, o gráfico normal demonstrou melhor desempenho, principalmente com a utilização do

procedimento proposto. Além disso, sob o modelo correto, as médias das medidas avaliadas parecem oscilar ao redor de um determinado patamar, que independe do tamanho amostral, da distribuição gerada ou do gráfico de probabilidade utilizado. Já dentre as duas formas de obtenção do envelope consideradas nesta análise, não há uma que se destaque em relação a outra.

Por fim, utilizamos para a aplicação do nosso método e das demais propostas os dados provenientes da Pesquisa Nacional de Saúde (PNS) de 2013, em que realizamos a modelagem da circunferência da cintura de indivíduos do Rio Grande do Sul em função de variáveis como idade e gênero, além de variáveis alimentícias. Através da aplicação dos procedimentos de rejeição a esses dados, constatamos que o nosso método não rejeitou a adequabilidade do modelo. Já os demais métodos, baseados nas propostas de Flack e Flores e Lee e Rhee, apresentaram decisão contrária, indicando a rejeição do ajuste.

7.1 Trabalhos futuros

Alguns trabalhos podem ser desenvolvidos a partir do procedimento de rejeição aqui apresentado. Um primeiro aspecto diz respeito ao estudo dos hiperparâmetros utilizados: 30; 50; 2,5 e 3,5. O valor 30 corresponde a quantidade de repetições do processo de obtenção dos limites do envelope fornecido ao usuário, enquanto o valor 50 refere-se ao número de réplicas utilizadas para a geração dos modelos supostos e seus respectivos envelopes. Os valores 2,5 e 3,5 são, respectivamente, os fatores multiplicativos associados a média da proporção de pontos fora e a distância entre a banda do envelope e o ponto mais afastado desse limite. O estudo de tais valores poderia ser conduzido para avaliar o efeito da alteração dos mesmos na performance do procedimento, para assim identificar uma variação que forneça resultados ainda melhores.

Outro ponto interessante se refere a escolha do resíduo tanto nos envelopes como no método de rejeição proposto. Este aspecto se mostra relevante, uma vez que para cada uma das classes de modelos abordadas utilizamos apenas uma forma residual nos estudos de simulação. Sendo assim, uma avaliação mais completa poderia ser conduzida no intuito de identificar se há resíduos com melhores características do que outros, levando em consideração os aspectos discutidos neste trabalho.

Apêndice A

Tabelas dos estudos de simulação para os modelos lineares generalizados

Tabela A.1: Resultados de simulação para o cenário II-a - Estatísticas descritivas para algumas medidas - Variável resposta com distribuição normal inversa.

Método	Medida	Gráfico	Normal						Meio normal					
		n	50		100		200		50		100		200	
		Distribuição ajustada	NI	GA	NI	GA	NI	GA	NI	GA	NI	GA	NI	GA
		Taxa rejeição proc. prop.	**0,056**	**0,359**	**0,052**	**0,608**	**0,067**	**0,919**	0,084	0,105	0,074	0,172	0,086	0,339
Atkinson	Prop. de vezes de pelo menos um ponto fora		0,864	0,971	0,943	0,996	0,966	1,000	0,834	0,853	0,919	0,927	0,957	0,972
	Prop. pontos fora	Média	**0,091**	**0,197**	**0,096**	**0,279**	**0,093**	**0,448**	0,100	0,107	0,108	0,109	0,104	0,132
		D.P.	0,085	0,139	0,085	0,157	0,082	0,163	0,108	0,111	0,116	0,108	0,107	0,125
	Maior dist. ponto envel.	Média	0,100	0,264	0,106	0,437	0,096	0,645	0,066	0,080	0,071	0,135	0,064	0,238
		D.P.	0,105	0,264	0,116	0,388	0,106	0,472	0,083	0,116	0,099	0,222	0,084	0,369
	Prop. pontos consec. fora	Média	**0,053**	**0,103**	**0,050**	**0,141**	**0,043**	**0,227**	0,062	0,067	0,060	0,061	0,053	0,066
		D.P.	0,049	0,084	0,047	0,102	0,043	0,126	0,070	0,076	0,070	0,068	0,060	0,073
Everitt/ Paula	Prop. de vezes de pelo menos um ponto fora		0,622	0,878	0,709	0,973	0,789	0,999	0,576	0,623	0,663	0,744	0,758	0,881
	Prop. pontos fora	Média	**0,042**	**0,133**	**0,042**	**0,204**	**0,043**	**0,393**	0,048	0,055	0,047	0,053	0,049	0,079
		D.P.	0,059	0,121	0,060	0,148	0,058	0,163	0,077	0,079	0,076	0,075	0,073	0,100
	Maior dist. ponto envel.	Média	0,050	0,191	0,054	0,348	0,051	0,579	0,034	0,042	0,036	0,084	0,033	0,191
		D.P.	0,086	0,231	0,093	0,337	0,079	0,452	0,072	0,091	0,079	0,182	0,063	0,338
	Prop. pontos consec. fora	Média	**0,029**	**0,077**	**0,026**	**0,108**	**0,024**	**0,211**	0,033	0,039	0,030	0,033	0,029	0,045
		D.P.	0,038	0,076	0,036	0,092	0,034	0,123	0,053	0,058	0,049	0,048	0,046	0,061

Tabela A.2: Resultados de simulação para o cenário II-b - Estatísticas descritivas para algumas medidas - Variável resposta com distribuição gama.

Método	Medida	Gráfico	Normal						Meio normal					
		n	50		100		200		50		100		200	
		Distribuição ajustada	GA	NI	GA	NI	GA	NI	GA	NI	GA	NI	GA	NI
		Taxa rejeição proc. prop.	0,047	0,414	0,054	0,782	**0,040**	**0,953**	0,072	0,477	0,078	0,811	0,067	0,934
	Prop. de vezes de pelo menos um ponto fora		0,897	0,962	0,941	0,998	0,976	1,000	0,839	0,943	0,903	0,993	0,950	0,999
Atkinson	Prop. pontos fora	Média	0,099	0,222	0,095	0,385	0,096	0,517	0,103	0,266	0,098	0,458	0,098	0,592
		D.P.	0,087	0,165	0,083	0,182	0,084	0,155	0,110	0,227	0,103	0,252	0,105	0,221
	Maior dist. ponto envel.	Média	0,116	0,401	0,113	0,718	0,103	1,055	0,071	0,368	0,068	0,685	0,063	1,021
		D.P.	0,119	0,485	0,129	0,706	0,109	0,914	0,097	0,493	0,097	0,717	0,086	0,918
	Prop. pontos consec. fora	Média	0,053	0,129	0,047	0,247	0,043	0,348	0,064	0,176	0,056	0,321	0,051	0,438
		D.P.	0,046	0,117	0,042	0,156	0,041	0,148	0,073	0,185	0,066	0,235	0,063	0,238
	Prop. de vezes de pelo menos um ponto fora		0,646	0,854	0,748	0,987	0,826	0,999	0,584	0,821	0,681	0,966	0,752	0,997
Everitt/ Paula	Prop. pontos fora	Média	0,048	0,157	0,047	0,312	0,047	0,470	0,048	0,195	0,049	0,378	**0,046**	**0,549**
		D.P.	0,063	0,152	0,060	0,180	0,060	0,162	0,076	0,214	0,073	0,257	0,072	0,230
	Maior dist. ponto envel.	Média	0,058	0,326	0,058	0,608	0,059	0,945	0,035	0,307	0,033	0,584	0,034	0,918
		D.P.	0,090	0,463	0,082	0,678	0,093	0,850	0,073	0,467	0,060	0,685	0,071	0,855
	Prop. pontos consec. fora	Média	0,031	0,100	0,028	0,213	0,025	0,336	0,034	0,138	0,032	0,273	0,027	0,424
		D.P.	0,038	0,110	0,035	0,151	0,031	0,150	0,054	0,176	0,049	0,234	0,045	0,240

Tabela A.3: Resultados de simulação para o cenário III-a - Estatísticas descritivas para algumas medidas - Variável resposta com distribuição normal inversa.

Método	Medida	Gráfico	Normal						Meio normal					
		n	50		100		200		50		100		200	
		Distribuição ajustada	NI	GA	NI	GA	NI	GA	NI	GA	NI	GA	NI	GA
		Taxa rejeição proc. prop.	**0,060**	**0,356**	**0,065**	**0,607**	**0,055**	**0,919**	0,080	0,099	0,084	0,170	0,081	0,360
	Prop. de vezes de pelo menos um ponto fora		0,870	0,967	0,936	0,994	0,969	1,000	0,831	0,848	0,911	0,926	0,940	0,977
Atkinson	Prop. pontos fora	Média	**0,090**	**0,206**	**0,096**	**0,269**	**0,089**	**0,448**	0,098	0,110	0,102	0,105	0,101	0,128
		D.P.	0,082	0,148	0,086	0,156	0,081	0,161	0,103	0,114	0,107	0,103	0,105	0,114
	Maior dist. ponto envel.	Média	0,100	0,285	0,107	0,438	0,101	0,744	0,069	0,090	0,066	0,139	0,067	0,312
		D.P.	0,109	0,293	0,120	0,400	0,120	0,537	0,091	0,138	0,093	0,250	0,101	0,434
	Prop. pontos consec. fora	Média	**0,052**	**0,110**	**0,049**	**0,134**	**0,042**	**0,233**	0,060	0,068	0,057	0,058	0,051	0,064
		D.P.	0,049	0,092	0,048	0,099	0,042	0,126	0,065	0,073	0,065	0,062	0,060	0,065
	Prop. de vezes de pelo menos um ponto fora		0,623	0,872	0,713	0,980	0,777	1,000	0,586	0,599	0,659	0,730	0,738	0,863
Everitt/ Paula	Prop. pontos fora	Média	**0,043**	**0,134**	**0,042**	**0,206**	**0,041**	**0,375**	0,048	0,054	0,047	0,055	0,047	0,069
		D.P.	0,059	0,122	0,057	0,141	0,057	0,165	0,077	0,081	0,073	0,079	0,075	0,085
	Maior dist. ponto envel.	Média	0,050	0,192	0,053	0,349	0,050	0,668	0,032	0,039	0,033	0,087	0,032	0,259
		D.P.	0,074	0,234	0,089	0,352	0,077	0,527	0,058	0,081	0,069	0,200	0,061	0,412
	Prop. pontos consec. fora	Média	**0,030**	**0,077**	**0,026**	**0,108**	**0,023**	**0,202**	0,034	0,038	0,030	0,036	0,028	0,040
		D.P.	0,040	0,075	0,035	0,088	0,033	0,121	0,054	0,057	0,048	0,054	0,047	0,052

Tabela A.4: Resultados de simulação para o cenário III-b - Estatísticas descritivas para algumas medidas - Variável resposta com distribuição gama.

Método	Medida	Gráfico	Normal						Meio normal					
		n	50		100		200		50		100		200	
		Distribuição ajustada	GA	NI	GA	NI	GA	NI	GA	NI	GA	NI	GA	NI
		Taxa rejeição proc. prop.	0,045	0,468	0,052	0,792	**0,061**	**0,978**	0,063	0,522	0,075	0,810	0,098	0,965
	Prop. de vezes de pelo menos um ponto fora		0,902	0,962	0,946	0,997	0,975	1,000	0,828	0,940	0,906	0,992	0,944	0,999
Atkinson	Prop. pontos fora	Média	0,096	0,221	0,096	0,386	0,099	0,555	0,099	0,270	0,099	0,460	0,102	0,659
		D.P.	0,087	0,167	0,082	0,181	0,083	0,140	0,105	0,231	0,103	0,247	0,106	0,198
	Maior dist. ponto envel.	Média	0,113	0,430	0,113	0,756	0,111	1,127	0,069	0,399	0,067	0,724	0,067	1,097
		D.P.	0,129	0,535	0,128	0,727	0,129	0,948	0,099	0,542	0,097	0,737	0,105	0,957
	Prop. pontos consec. fora	Média	0,053	0,130	0,048	0,244	0,044	0,399	0,061	0,177	0,055	0,320	0,053	0,509
		D.P.	0,046	0,118	0,044	0,154	0,041	0,138	0,068	0,185	0,064	0,233	0,062	0,233
	Prop. de vezes de pelo menos um ponto fora		0,660	0,867	0,743	0,985	0,823	1,000	0,587	0,820	0,671	0,973	0,745	0,997
Everitt/ Paula	Prop. pontos fora	Média	0,048	0,156	0,046	0,328	0,046	0,504	0,048	0,193	0,049	0,400	0,046	0,607
		D.P.	0,064	0,149	0,058	0,181	0,056	0,145	0,073	0,210	0,076	0,256	0,069	0,213
	Maior dist. ponto envel.	Média	0,058	0,326	0,061	0,644	0,057	1,008	0,033	0,306	0,034	0,622	0,033	0,985
		D.P.	0,090	0,452	0,095	0,677	0,081	0,878	0,073	0,457	0,065	0,683	0,063	0,882
	Prop. pontos consec. fora	Média	0,032	0,103	0,027	0,224	0,025	0,381	0,033	0,135	0,032	0,294	0,027	0,481
		D.P.	0,039	0,111	0,034	0,152	0,030	0,138	0,049	0,170	0,051	0,237	0,042	0,231

Tabela A.5: Resultados de simulação para o cenário IV-a - Estatísticas descritivas para algumas medidas - Variável resposta com distribuição normal inversa.

Método	Medida	Gráfico	Normal						Meio normal					
		n	50		100		200		50		100		200	
		Distribuição ajustada	NI	GA	NI	GA	NI	GA	NI	GA	NI	GA	NI	GA
		Taxa rejeição proc. prop.	**0,053**	**0,258**	**0,042**	**0,470**	**0,047**	**0,778**	0,081	0,093	0,060	0,173	0,077	0,316
	Prop. de vezes de pelo menos um ponto fora		0,870	0,957	0,940	0,992	0,974	0,998	0,833	0,844	0,911	0,924	0,951	0,958
Atkinson	Prop. pontos fora	Média	**0,091**	**0,171**	**0,094**	**0,220**	**0,096**	**0,322**	0,098	0,102	0,099	0,110	0,102	0,118
		D.P.	0,086	0,130	0,083	0,142	0,083	0,155	0,106	0,106	0,100	0,110	0,104	0,118
	Maior dist. ponto envel.	Média	0,104	0,242	0,106	0,405	0,102	0,607	0,069	0,083	0,066	0,153	0,068	0,261
		D.P.	0,113	0,266	0,113	0,405	0,117	0,524	0,095	0,134	0,082	0,269	0,099	0,423
	Prop. pontos consec. fora	Média	**0,051**	**0,091**	**0,047**	**0,107**	**0,044**	**0,154**	0,060	0,064	0,055	0,061	0,052	0,060
		D.P.	0,048	0,077	0,045	0,084	0,041	0,103	0,067	0,068	0,061	0,067	0,058	0,069
	Prop. de vezes de pelo menos um ponto fora		0,623	0,829	0,736	0,939	0,798	0,995	0,584	0,595	0,663	0,717	0,760	0,834
Everitt/ Paula	Prop. pontos fora	Média	**0,045**	**0,097**	**0,044**	**0,143**	**0,043**	**0,260**	0,050	0,049	0,047	0,052	0,048	0,070
		D.P.	0,061	0,103	0,057	0,121	0,056	0,151	0,077	0,074	0,072	0,074	0,074	0,095
	Maior dist. ponto envel.	Média	0,053	0,156	0,053	0,298	0,052	0,538	0,034	0,039	0,031	0,087	0,032	0,209
		D.P.	0,080	0,215	0,081	0,352	0,076	0,485	0,064	0,088	0,060	0,207	0,060	0,377
	Prop. pontos consec. fora	Média	**0,030**	**0,058**	**0,027**	**0,076**	**0,024**	**0,133**	0,035	0,035	0,031	0,033	0,028	0,041
		D.P.	0,040	0,063	0,034	0,072	0,032	0,100	0,054	0,052	0,047	0,049	0,046	0,059

Tabela A.6: Resultados de simulação para o cenário IV-b - Estatísticas descritivas para algumas medidas - Variável resposta com distribuição gama.

Método	Medida	Gráfico	Normal						Meio normal					
		n	50		100		200		50		100		200	
		Distribuição ajustada	GA	NI	GA	NI	GA	NI	GA	NI	GA	NI	GA	NI
		Taxa rejeição proc. prop.	0,035	0,298	0,040	0,579	0,051	0,856	0,072	0,360	0,069	0,625	0,076	0,868
Atkinson	Prop. de vezes de pelo menos um ponto fora		0,884	0,950	0,946	0,995	0,979	1,000	0,839	0,919	0,903	0,988	0,952	0,997
	Prop. pontos fora	Média	0,101	0,168	0,098	0,274	0,099	0,411	**0,100**	**0,194**	**0,097**	**0,328**	0,100	0,495
		D.P.	0,091	0,136	0,083	0,163	0,083	0,162	0,108	0,188	0,104	0,230	0,103	0,235
	Maior dist. ponto envel.	Média	0,115	0,340	0,114	0,503	0,108	0,754	0,069	0,296	0,067	0,455	0,068	0,704
		D.P.	0,123	0,442	0,123	0,515	0,125	0,697	0,091	0,443	0,091	0,511	0,099	0,695
	Prop. pontos consec. fora	Média	0,055	0,092	0,047	0,153	0,043	0,246	0,062	0,122	0,055	0,206	0,052	0,336
		D.P.	0,049	0,086	0,041	0,115	0,039	0,137	0,073	0,140	0,064	0,188	0,059	0,226
Everitt/ Paula	Prop. de vezes de pelo menos um ponto fora		0,650	0,805	0,769	0,956	0,824	0,998	0,566	0,746	0,689	0,924	0,749	0,992
	Prop. pontos fora	Média	0,046	0,105	0,049	0,205	0,048	0,353	0,047	0,127	0,049	0,255	0,049	0,437
		D.P.	0,060	0,117	0,063	0,161	0,060	0,161	0,075	0,162	0,073	0,228	0,075	0,234
	Maior dist. ponto envel.	Média	0,057	0,227	0,063	0,416	0,057	0,633	0,032	0,200	0,035	0,382	0,033	0,594
		D.P.	0,096	0,363	0,091	0,517	0,095	0,596	0,066	0,361	0,065	0,514	0,074	0,594
	Prop. pontos consec. fora	Média	0,030	0,064	0,028	0,126	0,026	0,226	0,033	0,085	0,031	0,170	0,029	0,308
		D.P.	0,036	0,076	0,036	0,114	0,032	0,131	0,053	0,122	0,046	0,185	0,046	0,217

Tabela A.7: Resultados de simulação para o cenário V-a - Estatísticas descritivas para algumas medidas - Variável resposta com distribuição normal inversa.

Método	Medida	Gráfico	Normal						Meio normal					
		n	50		100		200		50		100		200	
		Distribuição ajustada	NI	GA	NI	GA	NI	GA	NI	GA	NI	GA	NI	GA
		Taxa rejeição proc. prop.	**0,057**	**0,671**	**0,049**	**0,922**	**0,053**	**0,999**	0,080	0,234	0,062	0,391	0,072	0,685
Atkinson	Prop. de vezes de pelo menos um ponto fora		0,863	0,991	0,922	1,000	0,957	1,000	0,838	0,909	0,903	0,976	0,958	0,998
	Prop. pontos fora	Média	**0,088**	**0,322**	**0,088**	**0,465**	**0,091**	**0,645**	0,098	0,166	0,098	0,199	0,102	0,285
		D.P.	0,084	0,170	0,084	0,164	0,085	0,114	0,104	0,149	0,102	0,161	0,106	0,187
	Maior dist. ponto envel.	Média	0,092	0,390	0,093	0,633	0,097	0,941	0,068	0,110	0,067	0,199	0,068	0,379
		D.P.	0,101	0,309	0,096	0,440	0,112	0,572	0,092	0,126	0,085	0,276	0,100	0,476
	Prop. pontos consec. fora	Média	**0,051**	**0,173**	**0,047**	**0,254**	**0,043**	**0,395**	0,060	0,102	0,055	0,109	0,052	0,149
		D.P.	0,049	0,116	0,048	0,134	0,044	0,129	0,066	0,103	0,063	0,105	0,061	0,126
Everitt/ Paula	Prop. de vezes de pelo menos um ponto fora		0,566	0,963	0,691	1,000	0,766	1,000	0,578	0,737	0,676	0,877	0,740	0,980
	Prop. pontos fora	Média	**0,040**	**0,253**	**0,040**	**0,403**	**0,040**	**0,611**	0,049	0,102	0,047	0,135	0,047	0,209
		D.P.	0,060	0,163	0,058	0,167	0,057	0,114	0,078	0,124	0,073	0,142	0,072	0,172
	Maior dist. ponto envel.	Média	0,043	0,295	0,045	0,526	0,048	0,888	0,033	0,059	0,031	0,131	0,031	0,335
		D.P.	0,075	0,255	0,072	0,393	0,072	0,555	0,068	0,080	0,060	0,213	0,059	0,449
	Prop. pontos consec. fora	Média	**0,028**	**0,142**	**0,025**	**0,226**	**0,023**	**0,391**	0,034	0,069	0,032	0,082	0,027	0,116
		D.P.	0,041	0,108	0,037	0,131	0,035	0,123	0,053	0,089	0,051	0,095	0,045	0,114

Tabela A.8: Resultados de simulação para o cenário V-b - Estatísticas descritivas para algumas medidas - Variável resposta com distribuição gama.

Método	Medida	Gráfico	Normal						Meio normal					
		n	50		100		200		50		100		200	
		Distribuição ajustada	GA	NI	GA	NI	GA	NI	GA	NI	GA	NI	GA	NI
		Taxa rejeição proc. prop.	0,040	0,693	0,041	0,951	**0,052**	**0,999**	0,075	0,725	0,073	0,948	0,072	0,997
	Prop. de vezes de pelo menos um ponto fora		0,888	0,982	0,949	1,000	0,980	1,000	0,853	0,976	0,898	1,000	0,949	1,000
Atkinson	Prop. pontos fora	Média	0,098	0,352	0,095	0,554	0,095	0,708	0,100	0,438	0,098	0,656	0,101	0,819
		D.P.	0,089	0,205	0,081	0,168	0,081	0,102	0,106	0,269	0,102	0,225	0,105	0,127
	Maior dist. ponto envel.	Média	0,113	0,719	0,113	1,244	0,108	1,994	0,073	0,701	0,067	1,232	0,068	1,984
		D.P.	0,123	0,739	0,120	1,021	0,125	1,528	0,095	0,751	0,091	1,030	0,099	1,533
	Prop. pontos consec. fora	Média	0,054	0,235	0,047	0,416	0,042	0,570	0,063	0,321	0,054	0,533	0,051	0,726
		D.P.	0,049	0,177	0,041	0,174	0,038	0,107	0,072	0,255	0,062	0,256	0,060	0,169
	Prop. de vezes de pelo menos um ponto fora		0,650	0,932	0,760	0,996	0,818	1,000	0,567	0,917	0,687	0,994	0,764	1,000
Everitt/ Paula	Prop. pontos fora	Média	0,045	0,285	0,047	0,499	0,046	0,676	0,046	0,362	0,049	0,607	0,049	0,797
		D.P.	0,058	0,208	0,060	0,184	0,058	0,103	0,075	0,276	0,072	0,248	0,077	0,134
	Maior dist. ponto envel.	Média	0,057	0,626	0,060	1,149	0,056	1,866	0,034	0,616	0,035	1,141	0,034	1,859
		D.P.	0,095	0,724	0,088	1,053	0,094	1,388	0,070	0,731	0,064	1,059	0,074	1,392
	Prop. pontos consec. fora	Média	0,029	0,205	0,028	0,392	0,025	0,561	0,032	0,277	0,032	0,502	0,029	0,718
		D.P.	0,036	0,177	0,035	0,182	0,032	0,101	0,050	0,255	0,048	0,266	0,048	0,165

Tabela A.9: Resultados de simulação para o cenário VI-a - Estatísticas descritivas para algumas medidas - Variável resposta com distribuição normal inversa.

Método	Medida	Gráfico	Normal						Meio normal					
		n	50		100		200		50		100		200	
		Distribuição ajustada	NI	GA	NI	GA	NI	GA	NI	GA	NI	GA	NI	GA
		Taxa rejeição proc. prop.	**0,061**	**0,537**	**0,046**	**0,832**	**0,048**	**0,988**	0,080	0,167	0,065	0,281	0,072	0,504
	Prop. de vezes de pelo menos um ponto fora		0,864	0,982	0,930	0,999	0,970	1,000	0,835	0,892	0,914	0,956	0,953	0,991
Atkinson	Prop. pontos fora	Média	**0,089**	**0,270**	**0,089**	**0,389**	**0,093**	**0,560**	0,098	0,140	0,100	0,155	0,102	0,201
		D.P.	0,086	0,160	0,082	0,170	0,085	0,141	0,107	0,131	0,100	0,138	0,105	0,158
	Maior dist. ponto envel.	Média	0,098	0,357	0,098	0,540	0,101	0,784	0,069	0,105	0,068	0,176	0,068	0,299
		D.P.	0,105	0,319	0,100	0,417	0,122	0,526	0,093	0,137	0,085	0,265	0,102	0,434
	Prop. pontos consec. fora	Média	**0,052**	**0,143**	**0,047**	**0,198**	**0,044**	**0,310**	0,062	0,087	0,055	0,085	0,052	0,101
		D.P.	0,052	0,105	0,046	0,122	0,044	0,133	0,073	0,090	0,061	0,089	0,060	0,096
	Prop. de vezes de pelo menos um ponto fora		0,590	0,941	0,701	0,997	0,771	1,000	0,592	0,671	0,669	0,837	0,744	0,938
Everitt/ Paula	Prop. pontos fora	Média	**0,042**	**0,193**	**0,042**	**0,318**	**0,041**	**0,515**	0,050	0,077	0,047	0,096	0,047	0,134
		D.P.	0,061	0,147	0,057	0,165	0,057	0,142	0,079	0,103	0,072	0,114	0,073	0,137
	Maior dist. ponto envel.	Média	0,047	0,251	0,051	0,441	0,050	0,706	0,033	0,054	0,032	0,113	0,032	0,240
		D.P.	0,071	0,264	0,079	0,380	0,076	0,479	0,061	0,094	0,060	0,217	0,061	0,373
	Prop. pontos consec. fora	Média	**0,029**	**0,106**	**0,026**	**0,168**	**0,023**	**0,298**	0,035	0,051	0,031	0,059	0,028	0,074
		D.P.	0,041	0,092	0,036	0,116	0,033	0,130	0,056	0,071	0,049	0,074	0,045	0,085

Tabela A.10: Resultados de simulação para o cenário VI-b - Estatísticas descritivas para algumas medidas - Variável resposta com distribuição gama.

Método	Medida	Gráfico	Normal						Meio normal					
		n	50		100		200		50		100		200	
		Distribuição ajustada	GA	NI	GA	NI	GA	NI	GA	NI	GA	NI	GA	NI
		Taxa rejeição proc. prop.	0,046	0,437	0,041	0,736	**0,047**	**0,952**	0,072	0,492	0,075	0,747	0,072	0,947
	Prop. de vezes de pelo menos um ponto fora		0,886	0,959	0,947	0,995	0,979	1,000	0,835	0,940	0,896	0,989	0,943	0,999
Atkinson	Prop. pontos fora	Média	0,099	0,212	0,097	0,346	0,096	0,528	0,101	0,235	0,098	0,395	0,101	0,602
		D.P.	0,089	0,165	0,084	0,186	0,081	0,159	0,110	0,214	0,104	0,246	0,104	0,223
	Maior dist. ponto envel.	Média	0,116	0,423	0,114	0,739	0,107	1,188	0,071	0,390	0,067	0,702	0,067	1,156
		D.P.	0,125	0,518	0,122	0,729	0,122	1,052	0,095	0,526	0,091	0,735	0,097	1,058
	Prop. pontos consec. fora	Média	0,054	0,124	0,047	0,205	0,042	0,353	0,062	0,151	0,054	0,263	0,052	0,450
		D.P.	0,048	0,117	0,041	0,147	0,038	0,155	0,073	0,167	0,063	0,218	0,060	0,245
	Prop. de vezes de pelo menos um ponto fora		0,660	0,861	0,774	0,970	0,827	0,999	0,567	0,820	0,682	0,948	0,741	0,998
Everitt/ Paula	Prop. pontos fora	Média	0,046	0,148	0,048	0,280	0,047	0,480	0,046	0,168	0,049	0,328	0,049	0,560
		D.P.	0,058	0,150	0,061	0,192	0,059	0,166	0,074	0,193	0,074	0,255	0,077	0,228
	Maior dist. ponto envel.	Média	0,058	0,318	0,062	0,666	0,056	1,067	0,034	0,293	0,034	0,643	0,034	1,044
		D.P.	0,096	0,439	0,093	0,797	0,095	0,949	0,070	0,443	0,068	0,801	0,073	0,954
	Prop. pontos consec. fora	Média	0,030	0,094	0,028	0,181	0,026	0,340	0,032	0,115	0,031	0,232	0,029	0,435
		D.P.	0,035	0,108	0,035	0,152	0,032	0,155	0,051	0,153	0,048	0,222	0,047	0,236

Tabela A.11: Resultados de simulação para o cenário II - Estatísticas descritivas para algumas medidas - Variável resposta com distribuição poisson e binomial negativa.

Método	Medida	Gráfico	Normal						Meio normal					
		n	50		100		200		50		100		200	
		Distribuição gerada	PO	BN	PO	BN	PO	BN	PO	BN	PO	BN	PO	BN
		Taxa rejeição proc. prop.	0,079	0,502	0,069	0,705	0,063	0,926	0,119	0,549	0,100	0,774	0,084	0,954
	Prop. de vezes de pelo menos um ponto fora		0,849	0,966	0,916	0,994	0,946	1,000	0,729	0,929	0,830	0,987	0,884	1,000
Atkinson	Prop. pontos fora	Média	0,103	0,263	0,102	0,377	0,102	0,546	0,104	0,300	0,102	0,447	0,103	0,633
		D.P.	0,103	0,195	0,098	0,202	0,099	0,187	0,138	0,249	0,130	0,259	0,131	0,218
	Maior dist. ponto envel.	Média	0,124	0,378	0,121	0,453	0,113	0,528	0,077	0,286	0,072	0,346	0,068	0,415
		D.P.	0,146	0,326	0,135	0,317	0,121	0,315	0,113	0,276	0,104	0,263	0,091	0,263
	Prop. pontos consec. fora	Média	0,058	0,131	0,051	0,179	0,048	0,265	0,072	0,212	0,062	0,313	0,059	0,491
		D.P.	0,056	0,105	0,050	0,119	0,050	0,130	0,103	0,210	0,085	0,236	0,084	0,244
	Prop. de vezes de pelo menos um ponto fora		0,580	0,893	0,678	0,979	0,738	0,997	0,463	0,841	0,561	0,962	0,603	0,997
Everitt/ Paula	Prop. pontos fora	Média	0,049	0,194	0,049	0,306	0,048	0,483	0,050	0,231	0,052	0,382	**0,044**	**0,586**
		D.P.	0,071	0,178	0,073	0,205	0,069	0,202	0,099	0,232	0,100	0,260	0,087	0,241
	Maior dist. ponto envel.	Média	0,059	0,280	0,058	0,363	0,057	0,436	0,034	0,206	0,032	0,284	0,029	0,344
		D.P.	0,095	0,278	0,085	0,282	0,089	0,289	0,072	0,230	0,059	0,234	0,064	0,237
	Prop. pontos consec. fora	Média	0,032	0,103	0,030	0,151	0,026	0,246	0,036	0,168	0,035	0,278	0,027	0,472
		D.P.	0,044	0,098	0,042	0,114	0,037	0,135	0,073	0,194	0,069	0,236	0,055	0,255

Tabela A.12: Resultados de simulação para o cenário III - Estatísticas descritivas para algumas medidas - Variável resposta com distribuição poisson e binomial negativa.

Método	Medida	Gráfico	Normal						Meio normal					
		n	50		100		200		50		100		200	
		Distribuição gerada	PO	BN	PO	BN	PO	BN	PO	BN	PO	BN	PO	BN
		Taxa rejeição proc. prop.	0,072	0,492	0,069	0,707	0,063	0,939	0,101	0,575	0,107	0,782	0,104	0,964
Atkinson	Prop. de vezes de pelo menos um ponto fora		0,841	0,962	0,919	0,995	0,953	1,000	0,746	0,926	0,839	0,987	0,881	1,000
	Prop. pontos fora	Média	0,101	0,266	0,103	0,373	0,101	0,531	0,102	0,306	0,104	0,448	0,101	0,629
		D.P.	0,099	0,192	0,099	0,204	0,098	0,191	0,129	0,244	0,133	0,258	0,130	0,222
	Maior dist. ponto envel.	Média	0,125	0,383	0,123	0,457	0,111	0,526	0,077	0,296	0,073	0,363	0,066	0,426
		D.P.	0,144	0,322	0,134	0,323	0,116	0,304	0,118	0,281	0,108	0,274	0,091	0,252
	Prop. pontos consec. fora	Média	0,057	0,129	0,053	0,169	0,047	0,246	0,069	0,215	0,063	0,321	0,058	0,491
		D.P.	0,054	0,099	0,050	0,110	0,046	0,122	0,093	0,207	0,087	0,238	0,083	0,245
Everitt/ Paula	Prop. de vezes de pelo menos um ponto fora		0,580	0,914	0,688	0,977	0,740	0,999	0,460	0,855	0,556	0,965	0,648	0,997
	Prop. pontos fora	Média	0,047	0,199	0,045	0,299	0,048	0,479	**0,045**	**0,242**	0,046	0,384	0,049	0,594
		D.P.	0,069	0,177	0,065	0,203	0,069	0,199	0,087	0,232	0,091	0,260	0,093	0,231
	Maior dist. ponto envel.	Média	0,063	0,301	0,063	0,370	0,060	0,434	0,036	0,229	0,035	0,297	0,032	0,359
		D.P.	0,110	0,292	0,106	0,300	0,097	0,266	0,084	0,244	0,081	0,253	0,070	0,223
	Prop. pontos consec. fora	Média	0,031	0,104	0,027	0,142	0,027	0,231	0,033	0,175	0,032	0,279	0,031	0,477
		D.P.	0,041	0,094	0,036	0,107	0,037	0,125	0,063	0,194	0,066	0,234	0,061	0,247

Apêndice B

Tabelas dos estudos de simulação para os modelos de regressão para taxas e proporções

Tabela B.1: Resultados de simulação para o cenário II - Estatísticas descritivas para algumas medidas - Variável resposta com distribuição beta e beta retangular.

Método	Medida	Gráfico	Normal						Meio normal					
		n	50		100		200		50		100		200	
		Distribuição gerada	BE	BR	BE	BR	BE	BR	BE	BR	BE	BR	BE	BR
		Taxa rejeição proc. prop.	**0,043**	**0,176**	**0,041**	**0,361**	**0,054**	**0,665**	0,077	0,086	0,069	0,103	0,078	0,136
Atkinson	Prop. de vezes de pelo menos um ponto fora		0,888	0,934	0,955	0,9885	0,984	1,000	0,837	0,839	0,897	0,913	0,954	0,963
	Prop. pontos fora	Média	0,102	0,155	0,099	0,235	0,101	0,352	0,101	0,101	0,098	0,110	0,099	0,124
		D.P.	0,091	0,126	0,085	0,154	0,084	0,165	0,108	0,106	0,105	0,113	0,101	0,119
	Maior dist. ponto envel.	Média	0,110	0,151	0,111	0,177	0,107	0,211	0,065	0,073	0,064	0,077	0,066	0,079
		D.P.	0,119	0,149	0,121	0,143	0,124	0,152	0,089	0,110	0,084	0,113	0,098	0,122
	Prop. pontos consec. fora	Média	**0,056**	**0,084**	**0,048**	**0,118**	**0,044**	**0,182**	0,063	0,062	0,055	0,062	0,051	0,064
		D.P.	0,050	0,074	0,043	0,094	0,039	0,117	0,072	0,070	0,063	0,072	0,059	0,072
Everitt/ Paula	Prop. de vezes de pelo menos um ponto fora		0,664	0,790	0,779	0,926	0,830	0,990	0,554	0,585	0,713	0,701	0,742	0,810
	Prop. pontos fora	Média	0,047	0,096	**0,050**	**0,160**	0,049	0,279	0,047	0,049	0,052	0,059	0,049	0,067
		D.P.	0,061	0,106	0,063	0,135	0,060	0,161	0,076	0,074	0,079	0,086	0,074	0,092
	Maior dist. ponto envel.	Média	0,055	0,091	0,062	0,121	0,056	0,152	0,032	0,036	0,035	0,042	0,032	0,043
		D.P.	0,091	0,113	0,090	0,119	0,093	0,112	0,066	0,078	0,064	0,085	0,071	0,077
	Prop. pontos consec. fora	Média	**0,030**	**0,058**	**0,029**	**0,090**	**0,027**	**0,155**	0,033	0,035	0,033	0,037	0,029	0,040
		D.P.	0,037	0,067	0,037	0,086	0,033	0,109	0,053	0,053	0,053	0,055	0,047	0,062

Tabela B.2: Resultados de simulação para o cenário III - Estatísticas descritivas para algumas medidas - Variável resposta com distribuição beta e beta retangular.

Método	Medida	Gráfico	Normal						Meio normal					
		n	50		100		200		50		100		200	
		Distribuição gerada	BE	BR	BE	BR	BE	BR	BE	BR	BE	BR	BE	BR
		Taxa rejeição proc. prop.	0,037	0,437	0,037	0,814	0,054	0,986	0,077	0,454	0,069	0,850	0,074	0,995
	Prop. de vezes de pelo menos um ponto fora		0,892	0,991	0,949	1,000	0,979	1,000	0,836	0,976	0,905	0,997	0,950	1,000
Atkinson	Prop. pontos fora	Média	0,102	0,253	0,099	0,356	0,100	0,438	0,102	0,276	0,099	0,455	**0,098**	**0,643**
		D.P.	0,091	0,126	0,085	0,113	0,083	0,091	0,109	0,182	0,106	0,187	0,100	0,133
	Maior dist. ponto envel.	Média	0,112	0,218	0,111	0,330	0,107	0,475	0,065	0,127	0,064	0,166	0,066	0,194
		D.P.	0,118	0,123	0,122	0,146	0,123	0,144	0,086	0,090	0,086	0,084	0,097	0,073
	Prop. pontos consec. fora	Média	0,056	0,147	0,048	0,215	0,044	0,275	0,064	0,175	0,055	0,296	0,049	0,451
		D.P.	0,049	0,082	0,042	0,077	0,039	0,055	0,073	0,137	0,064	0,160	0,057	0,130
	Prop. de vezes de pelo menos um ponto fora		0,662	0,947	0,774	1,000	0,822	1,000	0,569	0,874	0,680	0,989	0,751	1,000
Everitt/ Paula	Prop. pontos fora	Média	0,047	0,192	0,050	0,304	0,048	0,399	0,048	0,204	0,050	0,395	0,048	0,603
		D.P.	0,059	0,119	0,063	0,105	0,060	0,082	0,077	0,176	0,077	0,186	0,073	0,137
	Maior dist. ponto envel.	Média	0,056	0,158	0,062	0,274	0,056	0,422	0,032	0,086	0,035	0,127	0,032	0,163
		D.P.	0,090	0,102	0,090	0,120	0,097	0,129	0,066	0,075	0,064	0,066	0,074	0,061
	Prop. pontos consec. fora	Média	0,030	0,125	0,029	0,201	0,026	0,267	0,035	0,140	0,032	0,270	0,029	0,436
		D.P.	0,036	0,081	0,036	0,073	0,034	0,052	0,057	0,134	0,051	0,156	0,047	0,126

Tabela B.3: Resultados de simulação para o cenário IV - Estatísticas descritivas para algumas medidas - Variável resposta com distribuição beta e beta retangular.

Método	Medida	Gráfico	Normal						Meio normal					
		n	50		100		200		50		100		200	
		Distribuição gerada	BE	BR	BE	BR	BE	BR	BE	BR	BE	BR	BE	BR
		Taxa rejeição proc. prop.	0,040	0,408	0,039	0,760	0,052	0,983	0,070	0,551	0,062	0,863	0,079	0,994
	Prop. de vezes de pelo menos um ponto fora		0,889	0,983	0,953	1,000	0,981	1,000	0,836	0,959	0,905	0,997	0,947	1,000
Atkinson	Prop. pontos fora	Média	0,102	0,252	0,098	0,382	0,101	0,530	**0,101**	**0,327**	0,099	0,512	**0,099**	**0,680**
		D.P.	0,092	0,149	0,085	0,147	0,084	0,117	0,108	0,225	0,106	0,209	0,101	0,138
	Maior dist. ponto envel.	Média	0,112	0,244	0,112	0,290	0,108	0,333	0,065	0,173	0,064	0,221	0,066	0,264
		D.P.	0,119	0,208	0,123	0,203	0,124	0,220	0,086	0,175	0,086	0,172	0,098	0,188
	Prop. pontos consec. fora	Média	0,055	0,139	0,048	0,198	0,043	0,265	0,063	0,227	0,056	0,387	0,050	0,556
		D.P.	0,049	0,091	0,042	0,088	0,038	0,068	0,072	0,188	0,066	0,202	0,059	0,146
	Prop. de vezes de pelo menos um ponto fora		0,659	0,922	0,773	0,997	0,827	1,000	0,546	0,885	0,691	0,982	0,753	1,000
Everitt/ Paula	Prop. pontos fora	Média	0,047	0,183	0,050	0,314	0,048	0,478	0,047	0,251	0,052	0,453	0,049	0,642
		D.P.	0,060	0,139	0,063	0,144	0,060	0,119	0,077	0,214	0,080	0,217	0,073	0,146
	Maior dist. ponto envel.	Média	0,056	0,168	0,062	0,219	0,056	0,263	0,032	0,118	0,035	0,168	0,032	0,212
		D.P.	0,092	0,158	0,091	0,167	0,096	0,170	0,066	0,130	0,064	0,144	0,074	0,144
	Prop. pontos consec. fora	Média	0,030	0,111	0,028	0,176	0,026	0,250	0,033	0,186	0,032	0,360	0,029	0,544
		D.P.	0,036	0,087	0,035	0,084	0,032	0,066	0,055	0,181	0,052	0,203	0,047	0,145

Tabela B.4: Resultados de simulação para o cenário V - Estatísticas descritivas para algumas medidas - Variável resposta com distribuição beta e beta retangular.

Método	Medida	Gráfico	Normal						Meio normal					
		n	50		100		200		50		100		200	
		Distribuição gerada	BE	BR	BE	BR	BE	BR	BE	BR	BE	BR	BE	BR
		Taxa rejeição proc. prop.	**0,051**	**0,556**	**0,056**	**0,821**	**0,046**	**0,992**	0,072	0,188	0,076	0,299	0,072	0,646
	Prop. de vezes de pelo menos um ponto fora		0,896	0,984	0,941	1,000	0,982	1,000	0,832	0,904	0,901	0,969	0,957	0,996
Atkinson	Prop. pontos fora	Média	0,099	0,301	**0,099**	**0,416**	**0,097**	**0,591**	0,099	0,159	0,101	0,200	0,098	0,321
		D.P.	0,085	0,163	0,084	0,157	0,079	0,116	0,106	0,141	0,105	0,147	0,102	0,159
	Maior dist. ponto envel.	Média	0,110	0,234	0,114	0,272	0,106	0,316	0,066	0,104	0,066	0,120	0,064	0,144
		D.P.	0,115	0,160	0,120	0,163	0,115	0,166	0,086	0,125	0,091	0,140	0,088	0,153
	Prop. pontos consec. fora	Média	**0,053**	**0,174**	**0,048**	**0,243**	**0,043**	**0,343**	0,062	0,101	0,057	0,118	0,051	0,184
		D.P.	0,046	0,108	0,042	0,109	0,037	0,078	0,069	0,103	0,065	0,104	0,060	0,111
	Prop. de vezes de pelo menos um ponto fora		0,667	0,945	0,752	0,996	0,826	1,000	0,587	0,731	0,663	0,871	0,739	0,977
Everitt/ Paula	Prop. pontos fora	Média	0,052	0,227	**0,046**	**0,358**	0,047	0,540	0,051	0,098	0,047	0,129	0,045	0,251
		D.P.	0,067	0,154	0,059	0,154	0,057	0,116	0,078	0,117	0,072	0,126	0,069	0,150
	Maior dist. ponto envel.	Média	0,058	0,171	0,059	0,218	0,056	0,261	0,033	0,064	0,034	0,079	0,032	0,102
		D.P.	0,083	0,133	0,093	0,140	0,081	0,127	0,059	0,100	0,071	0,118	0,058	0,122
	Prop. pontos consec. fora	Média	**0,033**	**0,145**	**0,027**	**0,227**	**0,025**	**0,331**	0,035	0,068	0,029	0,084	0,027	0,155
		D.P.	0,041	0,104	0,033	0,106	0,031	0,073	0,055	0,087	0,045	0,091	0,045	0,106

Tabela B.5: Resultados de simulação para o cenário VI - Estatísticas descritivas para algumas medidas - Variável resposta com distribuição beta e beta retangular.

Método	Medida	Gráfico	Normal						Meio normal					
		n	50		100		200		50		100		200	
		Distribuição gerada	BE	BR	BE	BR	BE	BR	BE	BR	BE	BR	BE	BR
		Taxa rejeição proc. prop.	**0,037**	**0,498**	**0,041**	**0,853**	**0,051**	**0,991**	0,075	0,192	0,065	0,363	0,072	0,636
	Prop. de vezes de pelo menos um ponto fora		0,891	0,979	0,953	0,998	0,983	1,000	0,834	0,884	0,897	0,970	0,948	0,9975
Atkinson	Prop. pontos fora	Média	0,102	0,277	0,099	0,436	0,101	0,578	0,102	0,144	0,098	0,221	0,099	0,328
		D.P.	0,090	0,166	0,085	0,159	0,084	0,119	0,108	0,136	0,106	0,161	0,100	0,174
	Maior dist. ponto envel.	Média	0,112	0,229	0,112	0,281	0,108	0,316	0,065	0,099	0,064	0,118	0,066	0,129
		D.P.	0,119	0,173	0,122	0,162	0,124	0,163	0,086	0,133	0,086	0,140	0,098	0,149
	Prop. pontos consec. fora	Média	**0,055**	**0,156**	**0,048**	**0,251**	**0,044**	**0,346**	0,063	0,091	0,055	0,132	0,050	0,194
		D.P.	0,048	0,109	0,042	0,115	0,039	0,086	0,071	0,095	0,065	0,113	0,058	0,129
	Prop. de vezes de pelo menos um ponto fora		0,658	0,925	0,772	0,995	0,833	1,000	0,553	0,699	0,695	0,877	0,748	0,973
Everitt/ Paula	Prop. pontos fora	Média	0,047	0,214	**0,050**	**0,375**	0,049	0,532	0,047	0,085	0,052	0,151	0,049	0,262
		D.P.	0,060	0,160	0,063	0,156	0,060	0,119	0,077	0,106	0,080	0,143	0,074	0,165
	Maior dist. ponto envel.	Média	0,056	0,165	0,062	0,225	0,056	0,266	0,032	0,059	0,035	0,078	0,032	0,091
		D.P.	0,093	0,142	0,091	0,139	0,095	0,124	0,066	0,100	0,064	0,112	0,073	0,107
	Prop. pontos consec. fora	Média	**0,030**	**0,131**	**0,029**	**0,232**	**0,026**	**0,336**	0,034	0,059	0,032	0,099	0,029	0,169
		D.P.	0,036	0,106	0,035	0,111	0,032	0,080	0,055	0,078	0,053	0,105	0,047	0,125

Referências Bibliográficas

ABESO, 2018. Associação Brasileira para o Estudo da Obesidade e da Síndrome Metabólica - Mapa da obesidade.
URL http://www.abeso.org.br/atitude-saudavel/mapa-obesidade 41

Anholeto, T., Sandoval, M. C., Botter, D. A., 2014. Adjusted pearson residuals in beta regression models. Journal of Statistical Computation and Simulation 84 (5), 999–1014. 36

Atkinson, A. C., 1981. Two graphical displays for outlying and influential observations in regression. Biometrika 68 (1), 13–20. 2, 13, 16, 17

Barros, M., Galea, M., Leiva, V., Santos-Neto, M., 2018. Generalized tobit models: diagnostics and application in econometrics. Journal of Applied Statistics 45 (1), 145–167. 2

Bayer, F. M., 2011. Modelagem e inferência em regressão beta. Ph.D. thesis, Universidade Federal de Pernambuco, Recife. 10

Bayer, F. M., Cribari-Neto, F., 2017. Model selection criteria in beta regression with varying dispersion. Communications in Statistics-Simulation and Computation 46 (1), 729–746. 9

Bayes, C. L., Bazán, J. L., García, C., et al., 2012. A new robust regression model for proportions. Bayesian Analysis 7 (4), 841–866. 8, 9, 10

Blom, G., 1958. Statistical estimates and transformed beta-variables. Ph.D. thesis, Almqvist & Wiksell. 14

Chernoff, H., Lieberman, G. J., 1954. Use of normal probability paper. Journal of the American Statistical Association 49 (268), 778–785. 14

Cook, R. D., Weisberg, S., 1982. Residuals and influence in regression. Chapman and Hall. 1, 5

Cordeiro, G. M., Demétrio, C. G., 2013. Modelos lineares generalizados e extensões. Piracicaba. 1

Cordeiro, G. M., Lima Neto, E. A., 2006. Modelos Paramétricos. Universidade Federal Rural de Pernambuco, Departamento de Estatística e Informática. 5

Daniel, C., 1959. Use of half-normal plots in interpreting factorial two-level experiments. Technometrics 1 (4), 311–341. 14

Davison, A., Gigli, A, 1989. Deviance residuals and normal scores plots. Biometrika 76 (2), 211–221. 6

Dunn, P. K., Smyth, G. K., 1996. Randomized quantile residuals. Journal of Computational and Graphical Statistics 5 (3), 236–244. 6, 31

Espinheira, P. L., Ferrari, S. L., Cribari-Neto, F., 2008. On beta regression residuals. Journal of Applied Statistics 35 (4), 407–419. 11, 12

Everitt, B. S., 1994. A Handbook of Statistical Analysis using S-Plus. Chapman and Hall. 17

Feng, C., Sadeghpour, A., Li, L., 2017. Randomized quantile residuals: an omnibus model diagnostic tool with unified reference distribution. arXiv preprint arXiv:1708.08527. 7

Ferrari, S., Cribari-Neto, F., 2004. Beta regression for modelling rates and proportions. Journal of Applied Statistics 31 (7), 799–815. 10, 11

Flack, V. F., Flores, R. A., 1989. Using simulated envelopes in the evaluation of normal probability plots of regression residuals. Technometrics 31 (2), 219–225. 2, 13, 17, 18, 24

Galea, M., Castro, M., 2017. Robust inference in a linear functional model with replications using the t distribution. Journal of Multivariate Analysis 160, 134–145. 2

Gan, F., Koehler, K., 1990. Goodness-of-fit tests based on p-p probability plots. Technometrics 32 (3), 289–303. 13

Green, P. J., 1984. Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. Journal of the Royal Statistical Society: Series B (Methodological) 46 (2), 149–170. 4

Hahn, E. D., 2008. Mixture densities for project management activity times: A robust approach to pert. European Journal of Operational Research 188 (2), 450–459. 8, 9

Hoaglin, D. C., Iglewicz, B., 1987. Fine-tuning some resistant rules for outlier labeling. Journal of the American Statistical Association 82 (400), 1147–1149. 18

Hoaglin, D. C., Iglewicz, B., Tukey, J. W., 1986. Performance of some resistant rules for outlier labeling. Journal of the American Statistical Association 81 (396), 991–999. 18

IBGE, 2018. Pesquisa Nacional de Saúde - Microdados.
URL ftp://ftp.ibge.gov.br/PNS/2013/microdados/pns_2013_microdados_2017_03_23.zip 41

Jones, M., 2009. Kumaraswamy's distribution: A beta-type distribution with some tractability advantages. Statistical Methodology 6 (1), 70–81. 8

Kieschnick, R., McCullough, B. D., 2003. Regression analysis of variates observed on (0, 1): percentages, proportions and fractions. Statistical modelling 3 (3), 193–213. 8

Klar, B., Meintanis, S. G., 2012. Specification tests for the response distribution in generalized linear models. Computational Statistics 27 (2), 251–267. 7

Kotz, S., Van Dorp, J. R., 2004. Beyond beta: other continuous families of distributions with bounded support and applications. World Scientific, Singapore. 8

Lee, J.-Y., Rhee, S.-W., 2001. Percentile envelope and its characteristic of error distribution for supernormality. Journal of the Korean Data and Information Science Society 12 (2), 35–45. 13, 19

Lemonte, A. J., Bazán, J. L., 2016. New class of johnson distributions and its associated regression model for rates and proportions. Biometrical Journal 58 (4), 727–746. 8

Lemonte, A. J., Bazán, J. L., 2018. New links for binary regression: an application to coca cultivation in peru. Test 27 (3), 597–617. 2

McCullagh, P., Nelder, J. A., 1989. Generalized linear models. Chapman and Hall, London. 4, 6, 14

MS, 2018. Ministério da Saúde - Só o IMC não diz como você está.
URL http://portalms.saude.gov.br/component/content/article/804-imc/40508 42

Nelder, J. A., Wedderburn, R. W., 1972. Generalized linear models. Journal of the Royal Statistical Society: Series A (General) 135 (3), 370–384. 3

Nóbrega, M. P., 2010. Estudo comparativo de gráficos de probabilidade normal para análise de experimentos fatoriais não replicados. Master's thesis, Universidade Federal do Rio Grande do Norte. 14

Olsson, U., 2002. Generalized linear models: an applied approach. Lund, Studentlitteratur, Sweden. 4

Paula, G. A., 2013. Modelos de regressão: com apoio computacional. IME-USP. 1, 4, 5, 17

Pereira, G. H., 2017. On quantile residuals in beta regression. Communications in Statistics-Simulation and Computation, 1–15. 1, 12, 15, 36

R Core Team, 2018. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
URL https://www.R-project.org/ 26

Raubertas, R. F., 1992. The envelope probability plot as a goodness-of-fit test. Communications in Statistics-Simulation and Computation 21 (1), 189–202. 2, 20

Ripley, B. D., 1977. Modelling spatial patterns. Journal of the Royal Statistical Society: Series B (Methodological) 39 (2), 172–192. 1, 15, 20

Santos, M. A. S., 2018. Representações Gráficas - Brasil Escola.
URL https://brasilescola.uol.com.br/fisica/representacoes-graficas.htm 1

Santos-Neto, M., Cysneiros, F. J. A., Leiva, V., Barros, M., et al., 2016. Reparameterized birnbaum-saunders regression models with varying precision. Electronic Journal of Statistics 10 (2), 2825–2855. 2

SBEM, 2018. Sociedade Brasileira de Endocrinologia e Metabologia - Números da Obesidade no Brasil.
URL http://www.endocrino.org.br/numeros-da-obesidade-no-brasil/ 41

Scudilio, J., Pereira, G. H., 2017. Adjusted quantile residual for generalized linear models. arXiv preprint arXiv:1710.11172. 6, 7, 27

Sen, P. K., Singer, J. M., De Lima, A. C. P., 2010. From finite sample to asymptotic methods in statistics. Vol. 28. Cambridge University Press, New York. 5, 10

Silva Ferreira, C., Vilca, F., Bolfarine, H., et al., 2018. Diagnostics analysis for skew-normal linear regression models: applications to a quality of life dataset. Brazilian Journal of Probability and Statistics 32 (3), 525–544. 2

Stasinopoulos, D. M., Rigby, R. A., et al., 2007. Generalized additive models for location scale and shape (gamlss) in r. Journal of Statistical Software 23 (7), 1–46. 9

Terrell, G. R., 2002. The gradient statistic. Computing Science and Statistics 34 (34), 206–215. 5

Wilk, M. B., Gnanadesikan, R., 1968. Probability plotting methods for the analysis for the analysis of data. Biometrika 55 (1), 1–17. 14

Willis, L. H., Slentz, C. A., Houmard, J. A., Johnson, J. L., Duscha, B. D., Aiken, L. B., Kraus, W. E., 2007. Minimal versus umbilical waist circumference measures as indicators of cardiovascular disease risk. Obesity 15 (3), 753–759. 42

Zahn, D. A., 1975. An empirical study of the half-normal plot. Technometrics 17 (2), 201–211. 14